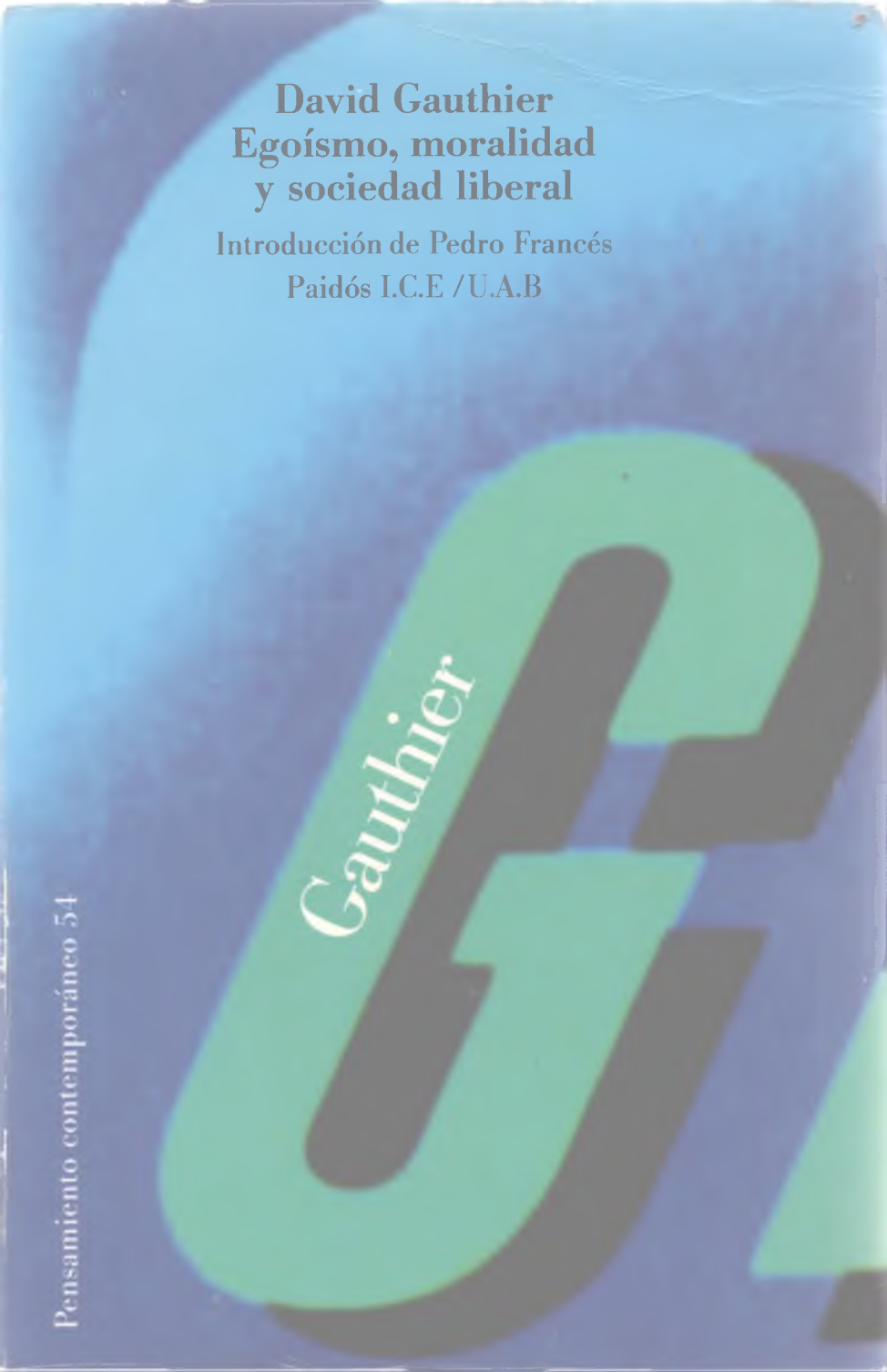
# David Gauthier
# Egoísmo, moralidad y sociedad liberal

Introducción de Pedro Francés

Paidós I.C.E /U.A.B




Pensamiento contemporáneo 54

# PENSAMIENTO CONTEMPORÁNEO
## Colección dirigida por Manuel Cruz

El objetivo de la colección *Pensamiento Contemporáneo* es proporcionar al lector interesado en este tema, y no sólo al especialista, un conjunto de textos de autores representativos del pensamiento de nuestro tiempo, textos en los que los mismos autores formulan de manera clara y concisa lo más significativo de su propuesta teórica, aquello que les ha convertido en clásicos de la filosofía del siglo XX.

6. T. S. Kuhn, *¿Qué son las revoluciones científicas?*
7. M. Foucault, *Tecnologías del yo*
8. N. Luhmann, *Sociedad y sistema: la ambición de la teoría*
9. J. Rawls, *Sobre las libertades*
10. G. Vattimo, *La sociedad transparente*
11. R. Rorty, *El giro lingüístico*
12. G. Colli, *El libro de nuestra crisis*
13. K.-O. Apel, *Teoría de la verdad y ética del discurso*
14. J. Elster, *Domar la suerte*
15. H.-G. Gadamer, *La actualidad de lo bello*
16. G. E. M. Anscombe, *Intención*
17. J. Habermas, *Escritos sobre moralidad y eticidad*
18. T. W. Adorno, *Actualidad de la filosofía*
19. T. Negri, *Fin de siglo*
20. D. Davidson, *Mente, mundo y acción*
21. E. Husserl, *Invitación a la fenomenología*
22. L. Wittgenstein, *Lecciones y conversaciones sobre estética, psicología y creencia religiosa*
23. R. Carnap, *Autobiografía intelectual*
24. N. Bobbio, *Igualdad y libertad*
25. G. E. Moore, *Ensayos éticos*
26. E. Levinas, *El Tiempo y el Otro*
27. W. Benjamin, *La metafísica de la juventud*
28. E. Jünger y M. Heidegger, *Acerca del nihilismo*
29. R. Dworkin, *Ética privada e igualitarismo político*
30. C. Taylor, *La ética de la autenticidad*
31. H. Putnam, *Las mil caras del realismo*
32. M. Blanchot, *El paso (no) más allá*
33. P. Winch, *Comprender una sociedad primitiva*
34. A. Koyré, *Pensar la ciencia*
35. J. Derrida, *El lenguaje y las instituciones filosóficas*
36. S. Weil, *Reflexiones sobre las causas de la libertad y de la opresión social*
37. P. F. Strawson, *Libertad y resentimiento*
38. H. Arendt, *De la historia a la acción*
39. G. Vattimo, *Más allá de la interpretación*
40. W. Benjamin, *Personajes alemanes*
41. G. Bataille, *Lo que entiendo por soberanía*
42. M. Foucault, *De lenguaje y literatura*
43. R. Koselleck y H.-G. Gadamer, *Historia y hermenéutica*
44. C. Geertz, *Los usos de la diversidad*
45. J. Habermas y J. Rawls, *Debate sobre el liberalismo político*
46. J.-P. Sartre, *Verdad y existencia*
47. A. Heller, *Una revisión de la teoría de las necesidades*
48. A. K. Sen, *Bienestar, justicia y mercado*
49. H. Arendt, *¿Qué es la política?*
50. K. R. Popper, *El cuerpo y la mente*
51. P. F. Strawson, *Análisis y metafísica*
52. K. Jaspers, *El problema de la culpa*
53. P. K. Feyerabend, *Ambigüedad y armonía*
54. D. Gauthier, *Egoísmo, moralidad y sociedad liberal*
55. R. Rorty, *Pragmatismo y política*

# David Gauthier

## Egoísmo, moralidad y sociedad liberal

Introducción de Pedro Francés

Ediciones Paidós
I.C.E. de la Universidad Autónoma de Barcelona
Barcelona - Buenos Aires - México

Títulos originales:
1. «Justice and Natural Endowment: Toward a Critique of Rawls's Ideologie Framework», publicado en *Social Theory and Practice*, 3, 1974, págs. 3-26.
© 1974 by Social theory and Practice
2. «The Incomplet Egoist», publicado en *Tannen Lectures on Human Values*, vol. 5, págs. 234-273.
© 1984 by University of Utah Press and Cambridge University Press
3. «Assure and Threaten», publicado en *Ethics*, 104 (julio de 1994), págs. 690-721.
© 1994 by The University of Chicago.
4. «Political Contractarianism», publicado en *Journal of Political Philosophy*, págs. 132-148.
© 1997 by Blackwell Publishers Ltd.

Traducción de Pedro Francés Gómez

Cubierta de Mario Eskenazi

*1ª edición, 1998*

Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del «Copyright», bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos.

© de todas las ediciones en castellano,
Ediciones Paidós Ibérica, S.A.,
Mariano Cubí, 92 - 08021 Barcelona, e
Instituto de Ciencias de la Educación
de la Universidad Autónoma de Barcelona,
08913 Bellaterra

ISBN: 84-493-0587-X
Depósito legal: B-18.000/1998

Impreso en Novagràfik, S.L.,
Puigcerdà, 127 - 08019 Barcelona

Impreso en España - Printed in Spain

# SUMARIO


INTRODUCCIÓN, *Pedro Francés* . . . . . 9

1. El origen del enfoque económico . . . . . 14
2. La moral por acuerdo . . . . . 20
3. Deliberación racional y teoría política . . . . . 26
4. Conclusión . . . . . 31

PREFACIO DEL AUTOR A LA EDICIÓN ESPAÑOLA . . . . . 35

1. Justicia y dotación natural: Hacia una crítica del marco ideológico de Rawls . . . . . 41
2. El egoísta incompleto . . . . . 67
3. Asegurar y amenazar . . . . . 119
4. Contractualismo político . . . . . 161

# Introducción

El egoísmo —entendido como la disposición a promover el interés propio, cualquiera que éste sea— es uno de los hallazgos conceptuales de la filosofía social, política y moral moderna. Según el contractualismo, el egoísmo es la argamasa de la sociedad y el fundamento último de la legitimidad del poder; y él solo basta, de acuerdo con los economistas clásicos, para explicar el orden y la eficiencia del mercado. Una sociedad de completos egoístas sería una sociedad económica y políticamente perfecta. Por otro lado, el gran esfuerzo del racionalismo no kantiano ha sido y es mostrar que las demandas morales pueden leerse como demandas del amor propio.[1] Así, una sociedad de completos, verdaderos egoístas ilustrados no sólo sería una sociedad perfecta, sino también una sociedad justa y feliz.

Tal vez algunos defensores de este modo de entender el proyecto moderno sostendrían que el bienestar que de hecho disfrutan las sociedades avanzadas de este fin de milenio se debe al relativo acercamiento a ese ideal, y que la causa de no haberlo realizado plenamente es que, debido a nuestras enraizadas supersticiones teológicas y morales, no sabemos ser completamente egoístas. Así, la ortodoxia económica sigue recomendando a particulares y gobiernos un egoísmo más perfecto: más previsor, más calculador, más resuelto. Y las teorías sociales, políticas, psicológicas (y hasta biológicas)[2] cuentan con el egoísmo como clave explicativa y predictiva. Pero, ¿hemos de fiarnos de tales recomendaciones y teorías? Desde luego, hay razones éticas para poner en cuarentena esta aclamación general del egoísmo pero, ¿y si hubiera también razones económicas?; ¿no sería sorpren-

1. Véase Savater, F., *Ética como amor propio*, Barcelona, Grijalbo, 1988.
2. Véase por ejemplo Dawkins, R., *El gen egoísta*, Barcelona, Salvat, 1997.

dente descubrir que el propio modelo de decisión y acción que el economista propone y exporta contradice hasta tal punto el egoísmo que lo hace analíticamente imposible?; ¿en qué lugar quedaría el ideal liberal moderno de sociedad si se probase que sólo podemos ser egoístas incompletos, y que el intento de «completar» nuestro egoísmo nos devuelve, paradójicamente, a la solidaridad y en último término a la moralidad?

Así puede interpretarse el desafío que plantea David Gauthier en los ensayos aquí reunidos; cuyo contenido gira alrededor de la tesis de que el egoísmo bien entendido contiene en sí mismo los recursos para explicar racionalmente ciertas conductas no egoístas, entre ellas la conducta moral. En cierto sentido, la propuesta de Gauthier sólo pretende explicar cómo es posible —si es que lo es— que el egoísmo sea de hecho el cemento de la sociedad y condición de la eficiencia del mercado, mostrando cómo puede construirse un argumento detallado que, partiendo del egoísmo, justifique racionalmente la cohesión social y explique la eficiencia económica.[3] Mas esta pretensión no es inocente. Plantea un desafío profundo, y de doble dirección: por un lado, cuestiona la validez del concepto dominante de racionalidad (el concepto «económico» o «instrumental», según el cual la razón es sólo el procedimiento para seleccionar los mejores medios, dados los fines del agente); pero ese cuestionamiento parte de la aceptación inicial de la racionalidad económica y sus premisas, lo que plantea, por otro lado, un desafío (metodológico, si se quiere) a la mayoría de los críticos de la razón instrumental.

Gauthier comparte con estos críticos la convicción de que una racionalidad exclusivamente instrumental está atrapada en graves limitaciones y contradicciones, pero su punto de vista «interno» desvía sus conclusiones, tanto del escepticismo o el nihilismo —destino final, por ejemplo, de algunos miembros de la escuela de Francfort—, como de la necesidad de buscar refugio en alguna forma alternativa de «racionalidad sustantiva» (como suele calificarse a los criterios valorativos y modelos de razonamiento práctico asociados a las creencias y tradiciones ideológicas reivindicadas por los teóri-

3. Un crítico ha escrito que lo que Gauthier pretende es «hacer visible la *mano invisible*» de la que habla Adam Smith. Véase Danielson, P., «The Visible Hand of Morality», *Canadian Journal of Pbilosopby*, vol. 18 (2), (junio de 1988), págs. 357-384.

cos «comunitaristas»). Las contradicciones de la razón instrumental exigen una reforma de la misma (ciertamente radical, según veremos), pero no necesariamente su abandono o su repudio. La de Gauthier es, por tanto, una crítica que no niega el valor de la racionalidad ilustrada —al menos en la versión que fue definida por Hume y los utilitaristas—, lo que le sitúa en el ámbito de discusión de la filosofía política y moral de raíz moderna (continental o anglosajona), que mantienen autores como Nozick, Dworkin, Habermas o Rawls. Ahora bien, la peculiaridad de la posición de Gauthier —su confianza en la posibilidad de una nueva comprensión de la razón instrumental— lo distancia tanto y en un sentido tan profundo de los autores citados que bien puede decirse que ellos son su punto de referencia polémico. Frente a Nozick y Dworkin, Gauthier defiende un fundamento contractual, no «natural» de los derechos e instituciones sociales; frente a Rawls, está convencido de la posibilidad de construir un verdadero contractualismo ético (y no meramente político); frente a Habermas, cree que los elementos para esa construcción pueden ser extraídos de los recursos de la racionalidad individual en su uso estratégico (sin necesidad de recurrir al uso comunicativo de la racionalidad).

Inscribir a Gauthier en el ámbito de la filosofía moral de raíz moderna sirve, como vemos, tanto para establecer el marco conceptual de su proyecto como para explicar su originalidad respecto a otros pensadores del mismo ámbito. En cuanto filósofo «moderno», Gauthier acepta la progresiva formalización de las demandas morales que se inició con Kant y cuyo último estadio es el llamado «procedimentalismo ético» (la tesis de que lo que define una norma como moral no es su contenido *ni su forma*, sino el hecho de haber sido producida mediante cierto procedimiento racional y, en consecuencia, lo que explica la universalidad de la ley moral no es ni el carácter de su contenido, *ni el hecho de carecer de contenido*, sino la posibilidad de que el procedimiento para su producción —y validación— sea universalmente aceptable). El procedimentalismo define las teorías de los autores antes citados. Ahora bien, las consecuencias que cada cual extrae de esta convicción común son muy diferentes. La posición que cuenta con más adeptos es quizá la que sostiene que, para ser consecuentes con el procedimentalismo, hay que limitar el papel normativo de la razón al ámbito (político) de creencias compartidas de una

sociedad concreta, pues éste es el único nivel en que cabe esperar un consenso universal sobre el procedimiento de validación de normas. Así pues, las normas pretendidamente válidas para todos los seres humanos (las normas morales) no pueden ser, estrictamente hablando, procedimentalmente evaluadas, y su único fundamento pueden ser las particulares creencias de los individuos o grupos. En consecuencia, las actitudes éticas basadas en tales creencias y normas morales estarían sujetas a discusión política y tendrían el reconocimiento o limitaciones que convencionalmente se establecieran pero, desde un punto de vista racional, no podrían ser consideradas válidas o inválidas. Parece, por tanto, que el intento procedimentalista de establecer el papel de la razón en materia práctica obliga a ceder a la razón (o sinrazón) individual el importantísimo campo normativo de la ética —para dejar en manos de la reflexión racional sólo el marco político y social donde la vida ética de cada cual va a desarrollarse—. El intento más interesante de superar esta limitación es el de la ética comunicativa, propuesta por Habermas. La ética comunicativa se basa en la constatación de que, a pesar del límite expuesto, que reduciría la capacidad normativa de la razón al establecimiento de un marco político para el desarrollo de cada individuo, las demandas éticas universalistas siguen estando presentes en la sociedad y reclaman su papel —a veces crítico, a veces constructivo— respecto a las normas que constituyen el marco político mencionado. Como quiera que parece imposible fundar procedimentalmente esas demandas en la razón instrumentalmente concebida, Habermas intenta hallar ese fundamento en una concepción comunicativa de la racionalidad. Sería muy prolijo, y fuera de lugar (además de una temeridad), intentar explicar en qué consiste la concepción comunicativa de la racionalidad y la ética derivada de ella. Conformémonos con la idea general siguiente: la razón humana tiene como una de sus funciones primordiales (no reductible a ninguna otra) la comunicación. Cuando la comunicación persigue un cierto objetivo —a saber, el acuerdo sobre las normas que afectan a *todos*— su misma estructura *exige* una cierta actitud en los participantes. Si esa actitud pudiera expresarse verbalmente, el contenido de esa expresión podría considerarse como una serie de mandatos obligatorios para los participantes en la discusión sobre normas, es decir, podría considerarse como el reglamento procedimental (universalmente válido) para discutir normas que afectan a to-

dos. Así, la propia naturaleza de la discusión sobre normas determinaría un único procedimiento válido para su legitimación, y ese procedimiento constituiría el principio de una ética comunicativa. Lo que nos interesa resaltar de este intento es su convicción de que la posibilidad de establecer un procedimiento universal (racional) para la validación de normas morales depende del concepto de racionalidad comunicativa, postulado como alternativa al uso instrumental (el más habitual) de la racionalidad, que es así totalmente descartado como fundamento de las normas que no sean meramente técnicas.

En resumen, el procedimentalismo parece conducir, bien a la retirada de la razón al ámbito de la política, cediendo a las creencias individuales el papel más importante en el debate ético; o bien a establecer una división en la razón, limitando la racionalidad instrumental al ámbito de la técnica e introduciendo la racionalidad comunicativa como fundamento de la normatividad, tanto política como, en última instancia, ética. La pretensión de Gauthier es no tener que hacer tales concesiones —ni la retirada al ámbito político, ni mucho menos el uso de una racionalidad irreductible al simple modelo instrumental—. Para Gauthier la propia racionalidad instrumental contiene, cuando es ejercida en interacción con otros seres racionales, la semilla de la actitud moral —que, aunque se traduce en imperativos eminentemente formales, también supone una limitación sobre los deseos o preferencias permisibles del agente—. Al desvelar cómo fructifica esa semilla, se muestra la racionalidad (y no la mera conveniencia) de la moralidad, pero también el carácter verdaderamente moral (y no meramente estratégico) de una racionalidad pragmáticamente efectiva para el logro de los fines del agente, de modo que no es necesario postular ningún otro modelo de racionalidad.

Esto convierte a Gauthier en un pensador original y polémico. Como teórico de la racionalidad se sitúa frente a la concepción ortodoxa de la racionalidad económica, y lo hace argumentando desde las propias premisas de ésta, en su propio terreno. Como filósofo moral, intenta demostrar la racionalidad de la moralidad sin desapegarse de la denostada visión instrumental que Hume declaró impotente fuera del reducido ámbito de los cálculos y las cuestiones de hecho. Como filósofo político, considera contractualmente justificada una sociedad civil liberal basada en el mercado como institución económica principal; un tipo de sociedad aproximadamente equidistante del

igualitarismo de Rawls y del radicalismo semilibertario de Nozick. Gauthier representa, por tanto, una posición peculiar, que añade a estos rasgos la audacia de emplear —con un nivel de especialización notable— el mismo lenguaje técnico del modelo económico de racionalidad cuya crítica y reformulación constituye el núcleo de su enfoque. Todo esto puede hacer que la lectura inicial de sus trabajos no sea un ejercicio fácil, y justifica las notas que siguen, cuya modesta pretensión es contextualizar brevemente los ensayos aquí reunidos, con la esperanza de ofrecer al lector algunas claves interpretativas que quizá no son explícitas en los textos.

## 1. *El origen del enfoque económico*

La investigación de la que estos ensayos forman parte tiene, según su autor, una concreta fecha de nacimiento: cierta tarde de noviembre de 1965, que Gauthier rememora al inicio de la segunda parte de «El egoísta incompleto». Aquella tarde alguien mencionó al joven profesor David Gauthier un juego bien conocido entre matemáticos y economistas: el Dilema del Prisionero. Este juego ilustra, de un modo sorprendentemente simple y preciso, el muy frecuente caso en que una disposición directamente interesada de todos los participantes en una interacción trae como consecuencia un resultado para cada uno (y, por ende, para todos) peor que el que podrían haber alcanzado adoptando cierta disposición altruista o «cooperativa». En un sentido más amplio, el Dilema del Prisionero evidencia la aptitud de la Teoría de Juegos como método para el análisis de los problemas de la interacción racional, entre los que se halla el problema de la justificación racional del comportamiento moral. El uso de tal metodología (entonces innovadora y hoy relativamente habitual, incluso entre filósofos) permitió a Gauthier formular con exactitud la cuestión que le obsesionaba desde el inicio de sus estudios: la reconciliación entre racionalidad y moralidad.

Gauthier había heredado el problema de R.M. Hare quien, compartiendo el credo analítico predominante en Oxford durante los años cincuenta y sesenta, se destacó por su intento de ofrecer una base racional para el lenguaje de la ética. La fidelidad al método analítico llevó a la mayoría de sus colegas a descartar la viabilidad de tal

empresa, y a condenar la filosofía práctica al emotivismo u otras formas de no cognoscitivismo. Sin embargo Hare, apoyado en la convicción de que «la respuesta a cuestiones morales es, o debería ser, una actividad racional»,[4] afirmó que «la tarea de la filosofía moral (...) es buscar un camino para reconciliar estas posiciones aparentemente incompatibles, resolviendo así la antinomia entre libertad y razón».[5] Como él afrontó esa tarea sin cuestionar un método basado exclusivamente en el análisis del lenguaje moral, su resultado se redujo a postular un criterio de decidibilidad racional para los juicios morales. Y algo parecido hizo Gauthier al principio, con resultado semejante.[6] Era evidente que el método analítico estaba seriamente limitado en su aplicación a la filosofía práctica, como claramente estableció Toulmin.[7] Incluso aunque aceptásemos como definitivo el criterio de Hare (el principio de universalizabilidad) para decidir sobre la validez de los juicios y argumentos morales, ese resultado estaría aún lejos del verdadero objetivo de la ética, que no es valorar locuciones, sino acciones; no (o no sólo) aclararnos qué prescripciones lógicamente válidas pueden considerarse morales, sino orientar efectivamente nuestras decisiones.

La Teoría de la Elección Racional,[8] empleada principalmente por los economistas, pretendía justamente eso: ser una orientación eficaz para la acción. Por ello la generalización de la misma que habían llevado a cabo matemáticos y teóricos de juegos pareció de inmediato —no sólo a Gauthier, desde luego—[9] el método más apropiado para una filosofía práctica capaz de retomar sus objetivos seculares, parcialmente abandonados varias décadas atrás.

4. Hare, R.M., *Freedom and Reason*, Oxford, Clarendon, 1963, pág. 2.
5. Ibid.
6. Véase Gauthier, D., *Practical Reasoning*, Oxford, Clarendon, 1963.
7. Véase Toulmin, S., *El puesto de la razón en la ética*, Madrid, Alianza, 1979.
8. La teoría de la decisión ha sido definida como «la colección de teorías matemáticas, lógicas y filosóficas que se ocupan de la toma de decisiones de los individuos —sea individualmente, en competencia o en grupos (...)—. Incluye la Teoría de la Utilidad, la Teoría de Juegos y la Teoría de la Elección Social» (Resnik, M.D., *Choices*, Minneapolis, University of Minnesota Press, 1993, pág. xi).
9. La primera vindicación del método económico para la filosofía moral fue quizá la de R.B. Braithwaite, *Theory of Games as a Tool for the Moral Philosopher*, Cambridge, Cambridge University Press, 1955. Esta obra tuvo una influencia decisiva en Gauthier y en toda una generación de filósofos morales.

La adopción del nuevo método permitió desplazar el centro de atención de la ética. En vez de los juicios morales y su validez, el objeto de investigación pasa a ser la acción y las razones para actuar; en vez de la argumentación moral y sus reglas, la decisión y acción moral (el altruismo) y su posible explicación racional. Y en este contexto, el problema fundamental de la filosofía moral es, como había detectado Kurt Baier —uno de los precursores de este enfoque—[10] la reconciliación del interés del agente con esa negación, o al menos restricción, del interés propio en la acción que suponen las normas morales.

El problema de la reconciliación entre las «razones del interés» y las «razones superiores», que corrigen o limitan las primeras en algunos casos, puede entenderse como una versión del problema de la reconciliación entre moralidad y racionalidad, determinada por la aceptación del método de la Teoría de la Decisión Racional. Este método permitía así una formulación simple y esquemática de un problema clásico. Pero la simplicidad tenía su precio: la formulación del problema en términos de la teoría de la decisión suponía admitir la prioridad de las «razones del interés». En efecto, la aceptación de la teoría normativa de la decisión (que pretende determinar cuál es la mejor decisión de un agente racional en un contexto dado) arrastró la de la teoría *descriptiva* de la decisión, cuyo objetivo es dar cuenta de los motivos reales de las decisiones de los agentes, pero que acaba, de hecho, introduciendo (sin duda dogmáticamente) el supuesto de que toda acción racional ha de poder ser descrita en función del interés (preferencia) de su autor. Pero la acción conforme a reglas (y en especial la acción moral o altruista) no encaja —más bien todo lo contrario— en ese tipo de descripción. De modo que, o bien se renuncia al concepto de racionalidad supuesto por la teoría descriptiva de la decisión (y con ello a la simplicidad de su planteamiento), o bien se idea un argumento que, respetando la prioridad del interés individual, sea capaz de justificar las limitaciones internas al mismo.

Afortunadamente, el esquema de ese argumento existía desde que, en los albores de la modernidad, fue necesario explicar la sumisión a las normas políticas sin apelar al origen divino del poder del soberano. Hobbes le dio forma al suponer que un sujeto político sólo aceptaría

10. Véase Baier, K., *The Moral Point of View*, Ithaca, (N.Y.), Cornell University Press, 1958.

voluntariamente convertirse en súbdito (cediendo al soberano la facultad de decidir por él cuestiones concernientes a su propia vida) si esperase obtener un beneficio para sí mismo superior al «coste» de esa cesión; y aun en ese caso, sólo si contase con que los demás miembros de la sociedad podrían aceptar (y aceptarían de hecho) las mismas reglas a las que él se somete. Con esa idea, Hobbes inició la tradición de lo que Macpherson ha denominado «teoría política del individualismo posesivo»[11] y a la vez —y sin sospecharlo siquiera— diseñó el único tipo de argumento capaz de explicar la necesidad de la unión social en términos comprensibles para individuos concebidos como «hombres económicos»: el contrato social moderno. Por eso no es de extrañar que el empleo del paradigma económico de racionalidad como herramienta metodológica en filosofía práctica fuese parejo con la recuperación del contractualismo como modelo de filosofía política y moral.

Gauthier fue uno de los iniciadores de esta recuperación con su obra *La lógica del Leviathan*,[12] que abrió un camino nuevo en la interpretación del contractualismo hobbesiano. La nueva ruta tiende a descreer las interpretaciones anteriores, que vieron en Hobbes el adalid del poder estatal absoluto y en su obra una justificación incluso de la tiranía. En su lugar, se impone la interpretación de que las conclusiones del *Leviathan* serían debidas (en caso de ser realmente tan radicales) a la ideología y el carácter de su autor —muy influido por su contexto social y político—, no al desarrollo lógico de sus premisas.[13] Si las premisas de Hobbes se analizan a la luz de la Teoría de Juegos, la monarquía absoluta es sólo uno de los resultados posibles de la interacción natural. Según ese análisis, un Estado constitucional, en que los súbditos retuvieran gran parte del poder y autonomía que gozan en el estado de naturaleza, sería un resultado más probable. De este modo, Hobbes es reintegrado con todos los honores a la tradición del contractualismo liberal, de donde había intentado ser desterrado.

11. Macpherson, C.B., *The Political Theory of Possessive Individualism. Hobbes to Locke*, Oxford, Oxford University Press, 1967.

12. Véase Gauthier, D., *The Logic of Leviathan*, Oxford, Clarendon, 1969.

13. Las contribuciones más importantes en esta línea han sido, sin duda, las obras de Jean Hampton, *Hobbes and the Social Contract Tradition*, Cambridge, Cambridge University Press, 1986, y de Gregory Kavka, *Hobbesian Moral and Political Theory*, Princeton (N.J.), Princeton University Press, 1986.

La fe en la posibilidad de reinterpretar y reconstruir el contractualismo hobbesiano no es compartida por otros pensadores neocontractualistas. Éste es el caso, eminentemente, de John Rawls. Y esta diferencia es crucial para entender la polémica entre Gauthier y él. La adhesión al modelo hobbesiano llevará a Gauthier, paradójicamente, a poder defender un «contractualismo moral», es decir, una teoría contractualista capaz de justificar la institución y mantenimiento no sólo de estructuras políticas y legales, sino también de restricciones y, sobre todo, disposiciones (modos de elección) personales que pueden denominarse «morales». El rechazo del modelo hobbesiano tendrá para Rawls la consecuencia contraria: el reconocimiento de que el contrato no genera ninguna consecuencia moral por sí mismo, de modo que la justificación contractual de las instituciones políticas ha de partir —para poder ser considerada justa— de un dato moral previo que no puede ser explicado en términos contractuales (lo que excluye de hecho, según dijimos arriba, la posibilidad de justificación racional). El filósofo político debe contar con que todas las personas poseen una «personalidad moral», expresada en su «sentido de la justicia» y en su propia «concepción del bien». Y tanto la idea de «personalidad moral» como las capacidades que la definen son previas al contrato, forman parte de las premisas que permiten que el resultado del contrato sea una sociedad democrática justa. El contractualismo sólo puede pretender ser político; la discusión ética —que conforma las personalidades morales de los ciudadanos— no le concierne. El problema de este enfoque es que, al concebir las opciones éticas como parte de las preferencias dadas de cada persona —de hecho, una de las preferencias básicas e irrenunciables—, no es necesario cuestionar la racionalidad formal (instrumental) que lleva a las partes a elegir una sociedad justa. Partiendo de un interés moral por la justicia suficientemente compartido, el uso instrumental de una razón neutra entre fines produce una sociedad justa; pero esto no tiene gran mérito.

Como es sabido, las premisas hobbesianas no mencionan la personalidad moral de las personas, y excluyen expresamente la existencia de derechos previos al contrato. Las partes se conciben, hipotéticamente, como agentes racionales (agentes económicos, diríamos hoy) en estado puro. Si el contrato tiene algún significado moral, si —como dice Gauthier— el contrato opera una «transformación mo-

ral»[14] en las personas, y si da lugar a una sociedad justa, todo ello ha de derivarse solamente de la racionalidad instrumental (o mera prudencia) de las partes, y de la estructura de interacción en que se hallan (y aquella en que esperan hallarse tras el contrato y la institución de la sociedad). Esto no es asunto fácil. Si aceptamos que las reglas de la prudencia se derivan de la experiencia y que las reglas morales son normas necesarias y universales, el tránsito de una a otra sería, según Kant, «una contradicción manifiesta», tan absurdo como pretender obtener agua de una piedra (*ex pumice aquam*).[15] Sin embargo, ésta es la tarea en que consiste, según Gauthier, un argumento ético «para adultos». Gauthier cree que, a pesar de las dificultades, es posible formular un argumento contractual que, partiendo de la simple (pero perfecta) prudencia de las partes y del exacto análisis de su situación y sus expectativas, demuestre que el comportamiento moral es racional. El argumento tiene, como ya ha quedado claro, la forma de un contrato social, es decir, la forma de un contrato hipotético entre *todos* los miembros de una sociedad. Pero lo más interesante es que su éxito coimplica una nueva comprensión de la racionalidad, cuyo concepto simplemente instrumental se revela así como una simple premisa metodológica, muy lejos de describir la *verdadera* naturaleza de la racionalidad humana. Al decir «coimplica» queremos dar a entender, primero, que el éxito del argumento demostraría que es racional abandonar la simple prudencia para abrazar una actitud moral sincera; y segundo, que no por ello habría que entender el argumento en un sentido lineal, como si el paso de la prudencia a la moralidad pudiera acontecer en el tiempo. El argumento mismo es posible porque la racionalidad instrumental resulta contradictoria en la práctica (su empleo consistente para lograr los fines del agente no siempre conduce a lograr los fines del agente) y la visión que nos dan de ella los teóricos «ortodoxos» no explica siempre el modo de decidir de los seres humanos, esto es, porque la racionalidad que estamos ya ejercitando de hecho (desde antes de la construcción del argumento contractual) no

14. Véase Gauthier, D., «Between Hobbes and Rawls», en Gauthier, D. y Sugden, R. (comps.), *Rationality, Justice and the Social Contract*, Ann Arbor, University of Michigan Press, 1993; págs. 24-39.

15. Véase Kant, I., *Crítica de la Razón Práctica*, Madrid, Librería General de Victoriano Suárez, 1913; en esp. prólogo, págs. 26 y 27.

es la descrita por la teoría ortodoxa y, por así decir, se las arregla para imponerse y hacerse evidente a través de esas contradicciones y paradojas que conducen al argumento moral. La posibilidad misma de una teoría ética para adultos revela que nuestra racionalidad es ya de hecho como la teoría moral sugiere que debe ser, y no como la teoría económica la describe.

Esto significa que el contractualismo moral, que decíamos llega parejo con el método económico de análisis de la acción, revela, no sólo que las normas morales están justificadas como guía de la acción racional, sino también que el concepto económico de racionalidad es (en su dimensión descriptiva) insatisfactorio. Así, el proyecto contractualista lleva consigo (y necesita) una teoría de la deliberación racional que supere las inadecuaciones de la teoría ortodoxa (identificada con la Teoría Bayesiana de la Decisión). Gauthier percibió esto y reconoció que sólo abriendo un frente doble —por un lado reformular el contractualismo hobbesiano para ofrecer una teoría contractual de la moralidad, y por otro analizar las contradicciones de la racionalidad instrumental para ofrecer una versión satisfactoria de la deliberación— sería posible realizar el anhelo de reconciliar la moralidad con la racionalidad. Así se perfilaron los dos grandes proyectos entrelazados que articulan su obra —y que de hecho están al servicio del anhelo filosófico de ofrecer un fundamento racional para la moralidad.

## 2. *La moral por acuerdo*

Al comienzo Gauthier no desarrolló una teoría global que articulase estos proyectos. Más bien al contrario. Las reflexiones sobre el contractualismo permanecieron en el nivel filosófico político, en el que el argumento depende lógicamente de la ideología (y ética) social imperantes —de ello es muestra «Justicia y dotación natural»—. Las reflexiones sobre la racionalidad se agruparon en algunos ensayos críticos sobre las nociones de «egoísmo», «coordinación» y «maximización»;[16] trabajos

16. Véase por ejemplo, «The Impossibility of Rational Egoism», *Journal of Philosophy*, vol 61, nº 14 (agosto de 1974), págs. 439-456; «Rational Cooperation», *Nous*, 8 (1974), págs. 53-65; «Reason and Maximization», *Canadian Journal of Philosophy*, 4 (1975), págs. 411-433; «Coordination», *Dialogue*, 14 (1975), págs. 195-221.

que sacaban partido a los dilemas y paradojas de la racionalidad prudencial, puestos de manifiesto (y de moda) por la Teoría de Juegos. Por último, el utillaje de esta teoría sirvió para formular la relación entre moralidad e interés con un lenguaje nuevo, pero esto produjo inicialmente conclusiones pesimistas, como si el nuevo método sólo hiciera más patente el abismo entre la prudencia y la moralidad, la imposibilidad de una moral racional.[17]

Lo más significativo de aquel período es que, frente al llamado «constructivismo kantiano» de Rawls, Gauthier había establecido una posición singular caracterizada, más que por sus demandas teórico-políticas liberales —ciertas, aunque muy dubitativamente expresadas entonces—, por el radicalismo en el análisis y por una fidelidad a las premisas del enfoque económico sólo comparable a la de economistas profesionales como Buchanan.

Con esta fidelidad y la inspiración hobbesiana como únicos puntos en común, Gauthier pensó reunir en un libro la serie de conexiones conceptuales que fue desarrollando durante un período de casi veinte años: conexiones entre maximización individual de la utilidad y optimización (o eficiencia) colectiva a través de la idea de «maximización individual restringida»; entre la libertad individual y las normas a través de su idea del «mercado perfecto como zona carente de moralidad»; entre las preferencias individuales y los valores sociales a través de la idea del contrato y del beneficio mutuo; y, por último, la conexión entre la racionalidad y la moralidad finalmente lograda a través de los conceptos de «negociación racional», «cooperación» y del «principio de concesión» como principio liberal de justicia distributiva.[18] Pero al recopilar estas ideas reconoció la posibilidad de articularlas en algo que hasta entonces sólo había acariciado, pero considerado imposible: una teoría contractual comprensiva que pudiera leerse como teoría moral, pero también como una contribución al neocontractualismo social y político en defensa de una sociedad liberal. A este período pertenece «El egoísta incompleto», que ya refleja la confianza en

17. Véase Gauthier, D. (comp.), *Morality and Rational Self-Interest*, Englewood Cliffs (N.J.), Prentice Hall, 1970, donde se incluye el ensayo «Morality and Advantage» (págs. 166-180), del que hay ed. española, en Raz, J. (comp.), *Razonamiento práctico*, México, FCE, 1986.

18. Véase sobre estos conceptos. Gauthier, D., *La moral por acuerdo*, Barcelona, Gedisa, 1994.

poder defender sin vacilaciones la racionalidad (prudencial) del comportamiento moral, empleando un argumento contractual según el cual los principios morales serían el resultado de una negociación hipotética entre agentes perfectamente racionales.

La idea maduró, y se extendió hasta incluir una original reivindicación contractual de los derechos individuales (que la tradición liberal había identificado con los «derechos naturales») como parte de las condiciones de estabilidad de un pacto social racional. El trabajo resultante fue la que quizá puede considerarse primera teoría moral contractualista en sentido estricto: *La moral por acuerdo.*

El esfuerzo y la audacia de Gauthier —no olvidemos que su proyecto seguía bajo el espectro del *ex pumice aquam* kantiano— se tradujeron en la repercusión inmediata de su obra, recibida como una de las contribuciones más relevantes al neocontractualismo inaugurado por Rawls. Así, James M. Buchanan, autor él mismo de una de las propuestas neocontractualistas más conocidas,[19] piensa que, «considerada como un gran argumento cuya conclusión es que las personas deberían (y posiblemente tienen que) adoptar la actitud moral incorporada en la estructura de Gauthier, la empresa logra ampliamente su objetivo».[20] Y algún comentarista ha llegado a afirmar que «la teoría del contrato social vuela, en *La moral por acuerdo*, más alto y con más pericia que nunca. Pero no vuela sola. Gauthier ha llevado a la misma altura de sofisticación el proyecto, a menudo cuestionado entre filósofos de todas las épocas, de deducir la moralidad a partir de la racionalidad (...). Nadie se había acercado tanto como Gauthier a llevar a cabo el proyecto con perfecto rigor y precisión».[21] Y, si estas afirmaciones parecen excesivas, lo que no se puede negar es el juicio de Robert Sugden, cuando recuerda que «entre las muchas contribuciones al proyecto contractualista, *La moral por acuerdo* de David Gaut-

19. Véase especialmente Buchanan, J., *The Limits of Liberty*, Chicago, University of Chicago Press, 1975; Véase también Buchanan, J. y Tullock, G., *The Calculus of Consent*, Ann Arbor, University of Michigan Press, 1965. Por sus trabajos sobre teoría económica de la democracia, J. Buchanan recibió el premio Nobel de Economía en 1986.

20. Buchanan, J.M., «The Gauthier Enterprise», *Social Philosophy and Policy*, 5:2, págs. 75-94, pág. 74.

21. Braybrooke, D., «Contract Theory's Fanciest Flight», *Ethics*, 97 (julio de 1987), págs. 750-764; pág. 751.

hier es probablemente la más significativa y, casi con seguridad, la más controvertida».[22]

Desde luego que la controversia y la crítica acompañaron a *La moral por acuerdo* desde su publicación. La audacia de sus planteamientos provocó una insistente revisión de cada una de las articulaciones teóricas de la propuesta. El hecho de tomar conceptos (y parte de su modo de análisis) de la economía facilitó que los expertos en estos temas recusaran algunas de las contribuciones más innovadoras de Gauthier —como su modelo de negociación racional, por ejemplo—. Aunque se trataba de problemas relativamente menores, estas críticas ponían en peligro la teoría como un todo, debido al difícil equilibrio que Gauthier tuvo que buscar para encajar todas las piezas de su proyecto. En este sentido, Buchanan —citado antes en su elogio de la teoría considerada como un gran argumento en favor de la aceptación de normas morales— cuestiona la validez del proyecto en cuanto intento de fundar un orden social liberal en el comportamiento racional (en sentido económico) de las personas. Un importante grupo de críticos cuestionó, desde una perspectiva filosófica, que Gauthier hubiera logrado reconstruir la moralidad con total fidelidad a su propósito de partir de premisas no morales. Calsamiglia habla de «mordiscos normativos» para referirse a distintas estipulaciones y precisiones que, de modo aparentemente inocuo, se van introduciendo en la teoría hasta llegar a constituir una verdadera premisa moral oculta, sin la cual no habría sido posible extraer un resultado con significado ético.[23]

Algunas de las críticas parciales lanzadas contra *La moral por acuerdo* quizá sean cuestionables porque parten de comprensiones asimismo parciales de la obra. Se podría recusar la visión económica que sólo ve en ella una justificación del papel de las normas estatales para asegurar la eficiencia que los mercados no pueden alcanzar por sí mismos, y una vaga alusión al fundamento ético de esas normas; o la visión ius-filosófica que percibe especialmente el error de intentar justificar racionalmente los órdenes normativos sociales y económicos sin

22. Sugden, R., «The Contractarian Enterprise», en Gauthier, D. y Sugden, R., *Rationality, Justice and the Social Contract*, *op. cit.*, págs. 1-23; pág. 1.

23. Véase Calsamiglia, A., «Un egoísta colectivo. Ensayo sobre el individualismo según Gauthier», *DOXA*, 6 (1989), págs. 77-94.

tener en cuenta las verdaderas motivaciones de la gente, que hacen a menudo necesaria la coacción; o la visión ética que lee la obra como un intento de pasar de la racionalidad prudencial a la racionalidad moral a través de puentes conceptuales no siempre resistentes a la crítica.

Pero lo cierto es que buena parte de las críticas apuntan a problemas reales de la teoría, y denuncian aspectos del proyecto que son sencillamente equivocados. Gauthier así lo ha reconocido.[24] La discusión provocada por *La moral por acuerdo* dejó claras, al menos, algunas cosas. En primer lugar, resulta patente que la Teoría de Juegos ha llegado a un grado de especialización y madurez tal que el filósofo moral puede contribuir muy poco a su desarrollo. Las modestas innovaciones que Gauthier intenta en su obra han quedado ampliamente superadas por los modelos axiomáticos propios de la teoría. Como mucho, el filósofo moral puede intentar casar sus intuiciones con los modelos formales de esta ciencia (y, desde luego, siempre queda, aunque ya fuera del terreno de la ética, el campo de la discusión metodológica). Por otro lado, los problemas básicos señalados por los críticos pueden condensarse en torno a la cuestión de la coherencia interna de la teoría: la dificultad se refiere a cómo mantener el relativismo axiológico característico de una sociedad liberal incorporado en la concepción instrumental de la racionalidad (y aceptado inicialmente por Gauthier como premisa de su argumento) mientras se insiste en el carácter moral —esto es, universal— de la normatividad que la razón impone en forma de restricciones al interés propio. No se trata de cuestionar ninguno de los elementos de la teoría, sino la posibilidad de reunirlos como Gauthier pretende. Otros esquemas que emplearan elementos similares podrían resultar coherentes: por ejemplo la justificación racional de un marco político y legal liberal en el que cupiese la libertad individual aceptada como punto de partida; o la justificación racional de una normatividad moral en la que se eliminase (o se aceptase sólo como hipótesis metodológica) el relativismo axiológico de las premisas, con la consecuencia de que el marco social subsiguiente no puede ser ya liberal, sino que para ser justo debe estar regido necesariamente por las normas mora-

24. Véase especialmente «Uniting Separate Persons», en Gauthier, D. y Sugden, R., *Rationality, Justice and the Social Contract*, *op. cit.*, págs. 176-192.

les que la teoría justifica; o incluso la aceptación de una «moral convencional», si es que tal término tiene sentido, en el marco de una sociedad y un Estado fundados instrumentalmente en el interés de sus miembros. Estas alternativas no satisfarían a Gauthier, desde luego, pero ejemplifican la coherencia de la que su propuesta parece carecer. El hecho de no aceptarlas como interpretación de su proyecto —bien insistiendo en el alcance moral, y no sólo político, de su contractualismo, bien admitiendo el relativismo axiológico y el marco liberal poscontractual, bien negando el carácter de «convencional» a la moralidad resultante del pacto— le enfrenta con el cargo general de incoherencia.

Gauthier, se resiste, sin embargo —y pese a haber admitido y rectificado llanamente las equivocaciones y oscuridades de su obra—, a renunciar al sentido global de su contractualismo moral. Y con la intención de restaurar las posibles fisuras, y llegar a persuadirnos de la viabilidad y racionalidad de su concepción, ha continuado trabajando sobre los problemas que comentaremos a continuación.

Antes es obligado precisar que las críticas que hemos «condensado» en esta dificultad genérica de coherencia no agotan, por supuesto, el espectro de las discusiones y polémicas generadas por la teoría de Gauthier. Las críticas incluidas (un tanto injustamente, por la necesaria simplificación) en ese problema general de coherencia, son las críticas que pueden llamarse «internas» —aquellas que no cuestionan necesariamente el individualismo liberal y el constructivismo moral que caracterizan el enfoque de Gauthier—. Desde luego que pueden formularse (y se han formulado de hecho) otro tipo de críticas, precisamente contra esos fundamentos. No hay duda de que el alcance de las mismas es, en un sentido, mayor que el de las críticas «internas», pero lo cierto es que Gauthier nunca se ha sentido tentado por esa polémica. Desde su punto de vista, renunciar al paradigma individualista liberal como base de argumentación significa renunciar al suelo común que permite contemporáneamente la discusión filosófico-práctica en general. Y la «carga de la prueba» recae en quienes pretenden llevar la discusión fuera del ámbito de conceptos modernos que articulan nuestra comprensión de nosotros mismos y nuestra sociedad. Desde luego que esta posición es muy cuestionable, pero es de hecho la razón por la que Gauthier ha considerado únicamente las críticas que le permitirían maniobrar dentro de su paradigma ideológico.

### 3. *Deliberación racional y teoría política*

Hecha esta salvedad, veamos, pues, cómo afronta Gauthier la cuestión general sobre la coherencia de su propuesta. En primer lugar hay que decir que, a pesar de que explícitamente Gauthier no renuncia a las líneas generales del proyecto, sí se observa una cesión de hecho, en el sentido de que su trabajo posterior a *La moral por acuerdo* se bifurca, abandonando en parte el objetivo de construir una teoría comprensiva. Por un lado, sigue la vía de la reflexión política o jurídica liberal que da por sentada la autonomía normativa (racional y moral) del individuo, y su prioridad respecto a la normatividad social. Por otro, profundiza notablemente en el desarrollo de una teoría de la deliberación racional capaz de explicar los casos (técnicamente denominados de «decisión dinámica»)[25] que escapan a la explicación ortodoxa porque no responden al esquema maximizador simple (que exigiría elegir la acción que maximiza la utilidad esperada en cada ocasión en que hay que optar entre acciones alternativas). Quizá se trate de una bifurcación sólo metodológica, y debamos esperar una nueva obra que reúna los progresos de ambos frentes, pero lo cierto es que la relativa independencia de los trabajos que se ocupan de uno y otro enfoque parece denunciar, cuanto menos, una duda sobre la posibilidad de unificar la teoría de la decisión racional individual con una teoría normativa comprensiva (moral, legal y política). En cualquier caso, se impone una mirada más cercana a cada una de estas vías, hacia el final de las cuales se sitúan los dos últimos artículos recogidos en esta colección.

El cargo de incoherencia, en que hemos cifrado el contenido de las críticas internas más relevantes, tiene dos posibles lecturas, y las líneas de trabajo recientes de Gauthier podrían verse como respuesta a cada una de ellas. La primera lectura, individual, se centra en la imposibilidad de explicar que sea racional, para agentes maximizadores, adoptar una actitud moral. El problema no estaría en la construcción de restricciones colectivas entre agentes interesados (o egoístas), sino en el cambio de actitud necesario para que su cumplimiento llegue a ser

25. Véase McClennen, E.F., *Rationality and Dynamic Choice*, Cambridge, Cambridge University Press, 1990.

considerado racional por agentes así concebidos. Según esta posición no hay argumento capaz de convencer a un agente maximizador de que puede ser en alguna ocasión racional que, viendo lo que es mejor para él (en términos de utilidad), haga lo que es peor. Y si tal argumento no existe, la empresa de Gauthier, que lo necesita para justificar deliberativamente el cumplimiento de las restricciones acordadas (y definir la moralidad como la clase de actitud incorporada en la disposición a cumplirlas), es inviable.

La segunda lectura, colectiva, no problematiza, en principio, el cumplimiento del pacto; en vez de eso sugiere que con las premisas individualistas y relativistas aceptadas no es posible deducir el tipo de pacto que Gauthier pretende. Gauthier se esfuerza en mostrar que el resultado de un contrato original entre individuos concebidos hipotéticamente como perfecta e igualmente racionales sería lo que denomina una «unión social». La idea subyacente es que la sociedad resultante del pacto original reflejaría las preferencias básicas de todas las partes del contrato (los miembros de la sociedad) de tal modo que sería capaz de concitar la adhesión emocional o sentimental de todos los ciudadanos. La crítica sostendría, por el contrario, que incluso concediendo la posibilidad de una justificación racional de la sociedad y sus instituciones mediante el argumento contractualista, la sociedad resultante sería, en palabras de Gauthier, una «sociedad privada», y no una unión social. Una sociedad privada se caracteriza porque los individuos retienen toda la capacidad normativa y aceptan las instituciones sólo en cuanto medios instrumentalmente eficaces para la promoción de sus intereses privados. Las instituciones de una sociedad privada no requieren adhesión emocional por parte de los ciudadanos, porque su mantenimiento está garantizado por la conveniencia de todos, y cuando esa conveniencia las hace superfluas o costosas, dejan de estar justificadas. Un teórico liberal no debería tener nada que objetar a una concepción «privada» de la sociedad. Sin embargo, no es ésta la posición de Gauthier, que claramente la deplora en «Justicia y dotación natural» y tampoco llega a aceptarla en *La moral por acuerdo*, donde dedica un capítulo completo a explicar cómo se desarrollaría la adhesión de los ciudadanos a la sociedad liberal fruto del pacto. Pues bien, esa insistencia en la unión social resulta —siempre según este análisis— incoherente con las premisas liberales.

La posición de la crítica puede expresarse sintéticamente en la te-

sis de que la suma de normatividades individuales no forma una normatividad colectiva distinta, ni la suma de reflexiones prudenciales llega a formar una reflexión moral. Esto no niega la normatividad social ni la moralidad, pero sí la posibilidad de deducirlas de la racionalidad prudencial individual, como Gauthier pretende.

Si tuviéramos que relacionar esta crítica con algún punto débil de *La moral por acuerdo*, pensaríamos en el argumento que justifica la adopción individual de la denominada «maximización restringida» (que es el argumento que explica cómo podría una suma de deliberaciones prudenciales concluir en la adopción de un modo de deliberación moral), y en la explicación de la sociedad liberal expuesta en la tercera parte de la obra. Si el proyecto quiere mantenerse a flote, parece imperativo reformular esos puntos.

Respecto al primer punto débil —o respecto a la primera lectura del cargo de incoherencia—, el enfoque económico había conducido a una formulación eficaz por lo paradójica: el agente verdaderamente maximizador es el que no maximiza; o el objetivo maximizador se logra mejor mediante acciones no maximizadoras (indirectamente maximizadoras, por tanto). Pero esta formulación fue llevada hasta su extremo en *La moral por acuerdo* sin lograr convencer a muchas personas, de modo que su rendimiento diríase agotado y, por otro lado, los propios economistas y teóricos de la racionalidad han venido a reconocer las limitaciones del modelo maximizador simple;[26] de modo que este argumento parecía haber llegado a un punto muerto, sin que su conclusión fuera aceptada. En vista de esto, Gauthier adopta una orientación pragmática y de sentido común, e inicia un análisis de la deliberación desde esa perspectiva. Los elementos básicos de esta orientación son, por un lado, las intuiciones sobre «lo que tiene sentido hacer» desde el punto de vista del agente, dados sus fines o propósitos y, por otro, el aparato analítico recién desarrollado por E. McClennen en su formalización de la decisión dinámica. La decisión dinámica se refiere a la decisión sobre cursos de acción o planes a lo largo del tiempo, y a las relaciones entre los nodos o puntos de decisión que se suceden en ellos. Por extensión, la decisión di-

26. Véase por ejemplo, Simon, H.A., *Models of Thought*, New Haven, Yale University Press, 1979; Sen, A., *Collective Choice and Social Welfare*, San Francisco, Holden-Day, 1970.

námica se refiere a las promesas, amenazas o intenciones sobre acciones futuras y, en sentido heurístico, podría aplicarse a la propia elección del modo de deliberación. Con estos elementos, se da forma a un argumento que no busca ya mostrar las contradicciones de una racionalidad maximizadora simple (cuestión que queda para los propios economistas), sino poner en evidencia que existe un modelo de deliberación que, por un lado, es pragmáticamente superior —en términos de logro de los fines del agente— al del modelo económico ortodoxo y que, por otro, refleja mejor la racionalidad humana tal como es intuitivamente percibida.

Si el rendimiento del enfoque económico se agotó, el de este enfoque pragmático es cuestionable. Pero, ciertamente, ha permitido a Gauthier retomar la discusión sobre la racionalidad (en términos del propósito del agente) de acciones que no tienden a promover ese propósito, lo cual mantiene la esperanza de poder justificar la moralidad en términos de interés. Por el momento, se trata sólo de ladrillos en el edificio de una adecuada comprensión de la deliberación racional. Y así debe entenderse el minucioso esfuerzo representado por «Asegurar y amenazar».

Por lo que respecta a la reformulación del contractualismo, algunos trabajos recientes de Gauthier parecen escritos para dar la razón a los críticos, en vez de para defender los postulados anteriores de Gauthier.[27] Adoptando una perspectiva intelectual que recuerda a la de Dworkin (aunque es independiente de ella), Gauthier acentúa el hecho de que las instituciones de un Estado constitucional democrático sólo tienen el poder que los ciudadanos depositan en ellas y con los límites que los ciudadanos deciden. Este argumento, que viene exigido por una concepción política profundamente individualista —acorde, como sostiene la crítica, con las premisas conceptuales de Gauthier—, se desliza sin remedio hacia una concepción privada de la sociedad, que filosóficamente pretende negarse. La incoherencia debió aquí ser patente al propio Gauthier, quien la reconoce en el artículo «Contractualismo político», rectificando así parte de lo dicho

27. Véase especialmente «Constituting Democracy», en Copp, Hampton y Roemer, (comps.), *The Idea of Democracy*, Nueva York, Cambridge University Press, 1993, págs. 314-334; «Public Reason», *Social Philosophy and Policy*, 12:1 (invierno de 1995), págs. 19-42.

en «Razón pública» (publicado dos años antes). Este problema traslu-ce lo que quizás es un conflicto irresuelto en el pensamiento de Gauthier: su convicción liberal en el terreno político y su fidelidad al argumento contractualista con un componente rousseauniano implícito muy fuerte. El argumento contractualista posee, pese a su raíz liberal, una lógica interna que lo aleja sorprendentemente del liberalismo, al menos del liberalismo político tal y como se entiende habitualmente. La idea del contrato social constituye un argumento constructivo que parte de premisas contrafácticas, es decir, no pretende ser un argumento histórico, ni reflejar en sus premisas ninguna verdad ontológica sobre el ser humano. La consecuencia es que, aunque ofrece un sólido fundamento individualista para la sociedad y el Estado, se trata en última instancia de un fundamento racional, no convencional. Tal argumento no niega que la sociedad (incluso el Estado) sean la condición de posibilidad del desarrollo individual, y aun del desarrollo histórico e intelectual que ha permitido su fundamentación racional desde premisas liberales —de modo que el contractualismo no implica necesariamente una actitud política que defienda la prioridad de los derechos individuales—. Por otro lado, instituciones que ya poseen una firme justificación racional no parecen necesitar ser concebidas, ni siquiera teóricamente, como fruto de la convención y el acuerdo entre individuos reales. Ésta es la idea que subyace a la defensa «contractual» de la institución de la familia, que podemos leer en «Contractualismo político». Ahora bien, esta consecuencia de un contractualismo radical (que Gauthier mismo califica como «revolucionario») choca con la ideología política liberal, en la medida en que tal ideología quiera hacerse pasar por el único modo racionalmente legítimo de entender la sociedad y el Estado. Este conflicto se resolvería si se reconociese que ciertas ideas políticas particulares —sobre la estructura del Estado o las relaciones entre derechos individuales y derechos colectivos— no son conclusiones necesarias de ninguna teoría pretendidamente neutral, sino fruto de una determinada (y legítima) ideología cuyo lugar es el ámbito de la discusión política pública, es decir, la zona de lo políticamente contingente dentro del gran marco de una sociedad contractualmente justificada. Con esto la incoherencia quedaría eliminada: la fidelidad a una ideología liberal no tendría por qué obstaculizar la defensa racional del tipo de sociedad lógicamente resultante del contrato.

Pero, como decíamos, Gauthier no es suficientemente claro a este respecto. No es fácil, a partir de sus textos, concluir hasta qué punto favorece una visión puramente instrumental de la sociedad, o más bien se inclina por la unión social que reclamó anteriormente. En esta oscuridad juega un papel importante el hecho de que las influyentes posiciones comunitaristas han obligado a Gauthier a optar decididamente por una sociedad liberal en sentido amplio, cuya prioridad sean los derechos y posibilidades individuales frente a los grupos y frente al Estado. Esta —quizá coyuntural— posición intelectual podría explicar la resistencia de Gauthier a reconocer explícitamente su deuda con Rousseau, y a resolver con ello, la incoherencia entre su concreto método contractualista y su posición ideológica liberal.

## 4. *Conclusión*

En conjunto, los textos seleccionados por Gauthier para este volumen pretenden señalar algunos hitos importantes de su camino intelectual y fijar su posición actual sobre los dos problemas teóricos que le preocupan: la teoría de la deliberación racional y el contractualismo como modelo de legitimación normativa.

El objetivo no puede ser más modesto: mostrar el difícil, pero sugerente, camino de este peculiar enfoque de la filosofía práctica que, empleando los métodos analíticos que han mostrado sus posibilidades en otras ciencias sociales (en particular, en economía), pretende a veces criticar, y siempre cuestionar, la base ideológica —cifrada en una particular concepción del individuo, la racionalidad individual y las relaciones entre las personas— de la sociedad que ha creado y fomentado esas mismas ciencias. Se trata del enfoque de la filosofía práctica más fácilmente asimilable por teóricos de otras áreas. De hecho, es el enfoque que mejor ha logrado, durante las últimas décadas, llevar la discusión filosófica y ética al campo de la economía, la política o el derecho. Se ha constituido así, en torno a las teorías de la deliberación y decisión racional, un nudo interdisciplinar al que el filósofo no puede permanecer ajeno.

El interés filosófico de estos textos no se ciñe, por otro lado, a la especialidad de un enfoque interdisciplinar o un método concreto, sino que abarca cuestiones clásicas —como la justificación racional

de la obligación moral, o la naturaleza de la deliberación prudencial— y se sitúa en el centro de uno de los debates filosófico-políticos más relevantes en la actualidad: el debate sobre la forma del liberalismo en sociedades multiculturales avanzadas.

Si el estilo en que estos debates están expresados hace quizá difícil el acceso a los mismos, también es verdad que permite una formulación potente y precisa de los postulados teóricos (éticos y políticos) de su autor. Los postulados que le llevan, en conclusión, a reivindicar, en un mundo de egoístas incompletos, un lugar privilegiado para el único egoísta completo: el agente moral.

PEDRO FRANCÉS
Universidad Complutense

## Principales obras de David Gauthier

*Practical Reasoning. The Structure and Foundations of Prudential and Moral Arguments and their Exemplification in Discourse*, Oxford, Clarendon, 1963.

*The Logic of Leviathan. The Moral and Political Theory of Thomas Hobbes*, Oxford, Clarendon Press, 1969 (1ª reimpresión, 1979).

*Morality and Rational Self-Interest* (comp.), Englewood Cliffs (Nueva Jersey), Prentice Hall, 1970.

*Morals by Agreement*, Oxford, Clarendon, 1986. Reimpreso en 1987 y 1988.

*Moral Dealing. Contract, Ethics and Reason*, Ithaca, Cornell University Press, 1990.

«"Le Promeneur Solitaire": Rousseau and the Emergence of the Post-Social Self», en *Social Philosophy and Policy*, otoño de 1990, págs. 35-58.

«Constituting Democracy», en D. Copp, J. Hampton y J.E. Roemer (comps.), *The Idea of Democracy*, Nueva York, Cambridge University Press, 1993, págs. 314-334.

*Rationality, Justice and the Social Contract. Themes from Morals by Agreement* (comp. con Robert Sugden), Ann Arbor, University of Michigan Press, 1993.

«Public Reason», en *Social Philosophy and Policy*, 12: 1 (invierno de 1995), págs. 19-42.

«Intentions and Deliberation», en Peter Danielson (comp.), *Modelling Rationality, Morality and Evolution*, Oxford, Oxford University Press, 1998.

## Traducciones al castellano

«La moral y la ventaja», en Joseph Raz (comp.), *Razonamiento práctico*, México, FCE., 1986 (trad. de Juan José Utrilla).

«¿Por qué contractualismo?», *DOXA*, 6 (1989), págs. 19-38, (trad. de Silvia Mendlewicz y A. Calsamiglia).

*La moral por acuerdo*, Barcelona, Gedisa, 1994 (trad. de Alcira Bixio).

# PREFACIO DEL AUTOR A LA EDICIÓN ESPAÑOLA

El primero de los ensayos traducidos para este volumen precede al último en más de veinte años. Poco puede sorprender que mis ideas hayan evolucionado a lo largo de tal período. Releer «Justicia y dotación natural» me resulta una experiencia un tanto inquietante. Por un lado, me encuentro perfectamente de acuerdo con mi deducción de los principios de justicia que Rawls debería haber derivado de su argumento contractualista. Pero me deja perplejo la última sección de mi artículo, donde trato la concepción de racionalidad empleada por Rawls como parte de la peculiar perspectiva ideológica de nuestra sociedad. En la mayor parte de mi trabajo reciente —ejemplificado aquí por «Asegurar y amenazar»— intento ofrecer una explicación preliminar de ciertos aspectos de una teoría de la deliberación racional, y los muy reales problemas que se plantean en mi intento no son ideológicos. Actualmente no creo que la racionalidad sea un concepto ideológico; una adecuada teoría de la racionalidad debería tener validez universal. Por otro lado, en aquella sección final de «Justicia y dotación natural», no me atrevo a dar una opinión definitiva sobre la aceptabilidad del individualismo liberal, que ahora apruebo sin vacilar. «Contractualismo político» contiene implícitamente mi respuesta positiva a Rawls, porque muestra qué normas y estructuras acordarían racionalmente las personas si —concebida la sociedad como una «empresa cooperativa para el beneficio mutuo»— tuvieran la oportunidad de decidir entre todas las condiciones de funcionamiento de tal empresa.

El tema principal de mis investigaciones filosóficas queda establecido en la primera frase de «Justicia y dotación natural»: «El problema principal de la Ética moderna es la reconciliación de la moralidad con la racionalidad». De esto sigo tan convencido como cuando es-

cribí esas palabras o como cuando, años antes —siendo estudiante posgraduado de filosofía—, reflexioné por primera vez sobre este problema. He buscado una solución que fuese convincente para cada individuo concreto, mostrándole por qué es racional aceptar restricciones morales al deliberar y, en consecuencia, al decidir y actuar. Y creo haber encontrado esa solución; aparece en la respuesta que doy a la pregunta «¿qué debe hacer un egoísta?», que inicia la segunda parte de «El egoísta incompleto». La clave de mi solución está en reconocer que hay beneficios comunes que pueden realizarse si se aceptan restricciones mutuas, de modo que es razonable para cada persona adoptar prácticas morales, tales como mantener las promesas, decir la verdad, ser honrado y practicar la justicia —y, en general, restringir la persecución de sus propios intereses tratando a los otros siempre como fines y no sólo como medios— siempre que se den dos condiciones. La primera es, por supuesto, que espere que los demás también cumplan con estas prácticas, al menos la mayoría de las veces, de modo que también ella será tratada como un fin. La segunda condición es que pueda esperar que su conformidad con las prácticas morales le permita acceder a oportunidades más beneficiosas que aquellas de las que disponen los egoístas, gracias a que su fiabilidad, honradez y preocupación por sus semejantes la conviertan en un socio aceptado en los sistemas de cooperación mutuamente ventajosos.

Ahora reconozco que este argumento tiene una limitación importante, en la medida en que se dirige únicamente a aquellas personas para quienes la interacción cooperativa con otros mejora las expectativas de realizar sus intereses. Pero sólo esas personas pueden ser bien recibidas por el resto de nosotros en una sociedad concebida como una empresa cooperativa. Si una persona tiene unos objetivos tales que no le dan ninguna razón para adoptar restricciones morales en su interacción con nosotros, entonces nosotros no tenemos ninguna razón para aceptar tales restricciones en nuestra relación con ellos. En mi opinión, la moralidad es esencialmente recíproca.

La reconciliación entre moralidad y racionalidad tiene implicaciones para nuestra comprensión de la moralidad y, como veremos enseguida, también para nuestra comprensión de la racionalidad. Mi reconciliación depende de considerar que las prácticas morales nos procuran un beneficio o ventaja mutua gracias a restricciones mutuas.

Pero no es que yo afirme que toda práctica mutuamente beneficiosa sea moral. Hasta la esclavitud podría ser mutuamente beneficiosa en comparación con el estado de anarquía resultante de la ausencia de toda restricción social y moral, pero no por ello está moralmente justificada. Aquí apelo a la idea de un acuerdo hipotético. Aquellas prácticas —e instituciones sociales que las incorporan— libremente consentidas por personas racionales hipotéticamente puestas en situación de decidir conjuntamente los términos de su interacción, en ausencia de coacción, y en tales circunstancias que ninguna se beneficiara de la presencia de otros de un modo que hiciera su presencia perjudicial para todos los demás, sólo esas prácticas están totalmente justificadas desde el punto de vista moral. Cuando el beneficio mutuo se logra mediante prácticas que serían racional y libremente acordadas por todos, y el respeto a las mismas será estable. Pueden darse sistemas de cooperación mutuamente ventajosos que permitan, sin embargo, una distribución muy desigual de esas ventajas, de modo que alguien se lleve «la parte del león». Los menos favorecidos por esos sistemas verán justificado el intento de modificarlos, de modo que, aunque respetaran las normas de los mismos, su conformidad no sería estable. El requisito de que las prácticas sean objeto de un acuerdo libre elimina esta fuente de inestabilidad. Cada persona las adopta y respeta para poder ser bienvenida como participante en un modo de vida social libremente elegido que funciona para el beneficio de todos.

Así, la reconciliación entre moralidad y racionalidad nos lleva a una moralidad contractualista. Y tiene también, como decía, implicaciones relativas a la racionalidad. Yo defiendo que es racional que las personas deliberen según ciertas restricciones referidas a las persecución de los propósitos (*aims*) del agente —aquellas que hubieran sido adecuadamente elegidas—. Sin embargo, la opinión ortodoxa, según la cual la deliberación racional selecciona la acción que maximiza el beneficio esperado del agente o, en general, la expectativa de lograr su propósito (*aim*), es incompatible con la demanda de que ciertas restricciones intervengan en la deliberación racional. Quizá sea beneficioso para ambos el intercambio en el que cada uno acepta la restricción de una promesa, pero para mí es aún más beneficioso que tú mantengas tu promesa mientras yo me libero de toda restricción y reniego de la mía. Por mi parte, sostengo que es con todo racional para mí mantener mis promesas, porque así puedo esperar beneficiarme de

las oportunidades que se presentan a las personas que tienen palabra.

Debo, por tanto, rechazar la concepción ortodoxa de la racionalidad maximizadora. Pero en ningún caso pretendo eliminar todo nexo entre la deliberación racional y el logro de los propósitos del agente. Al contrario, sostengo que los procedimientos racionales de deliberación son aquellos de los que se pueda esperar que conducirán mejor a la realización de los propósitos del agente, siempre que se empleen consistentemente. La deliberación racional se caracterizará, en consecuencia, por la honestidad y fiabilidad que convierte a alguien en un participante bienvenido a la cooperación, ya que, como he dicho, son estas personas las que tienen mayores perspectivas de lograr sus propósitos. Desde mi punto de vista, una razón para actuar es una consideración a favor o en contra de realizar cierta acción, que interviene en los procedimientos de deliberación racional. Así, el hecho de haber prometido hacer algo contará como una razón para hacerlo, independientemente de si esa realización contribuye o no a lograr mis propósitos, porque las promesas deben ser tratadas como una razón si quiero que mi modo de deliberación me confiera la confianza y fidelidad de otros.

Estos brevísimos comentarios pretenden únicamente situar los temas principales de mi trabajo y señalar cuál es la relación entre los mismos. Lo que he dicho habrá ya aclarado que cada uno de estos tres temas principales —la reconciliación de moralidad y racionalidad, la base contractualista de la moralidad y la naturaleza de la deliberación racional— está representado en esta colección. En «Justicia y dotación natural» acepto, con Rawls, la base contractualista de la justicia, aunque rechazo su versión de los principios que acordaríamos. En la segunda parte de «El egoísta incompleto» expongo el argumento que reconcilia moral y razón, mostrando por qué es racional aceptar restricciones morales. Muestro así que es irracional ser un egoísta puro —es decir, todo lo puro que se puede, ya que en la primera parte del artículo muestro que, supuesta cierta descripción razonable de lo que significa ser un egoísta, el egoísmo puro resulta imposible—. En «Asegurar y amenazar» describo parte de la teoría de la deliberación acorde con este argumento «reconciliador». Y en «Contractualismo político» retorno a la base contractual, no de la moralidad, sino de la sociedad en este caso, para esbozar los rasgos básicos de una sociedad que pudiera contar con el asentimiento libre

de todos, en una hipotética elección conjunta de los términos de nuestra interacción.

Aunque se aceptara todo lo que defiendo en estos artículos, ninguno de los proyectos que persigo en ellos está completo. He tratado de reconciliar la moralidad con la racionalidad desde una perspectiva hobbesiana, encontrando una gran analogía entre mis argumentos y los que emplea Hobbes en defensa de la justicia frente a su imaginario antagonista, el «Necio» (*the «Foole»*). Pero, como sugiere la mención anterior al trato de las personas como fines, mi argumento posee una dimensión kantiana. Hace algunos años escribí un artículo, «La unidad de la razón», que subtitulé «Una reinterpretación subversiva de Kant»;[1] y así es, sin duda, como debería verse la dimensión kantiana de mi trabajo. Al contrario que Kant, yo ignoro la «razón pura práctica» y me centro en la interacción racional. Pero en mi próximo artículo, «Los fines de la razón» (*«Reason's ends»*), defiendo que hay un lugar para el imperativo categórico en mi teoría. Creo que el error de Kant es que concibe la moralidad y la racionalidad en tan íntima conexión, que ni siquiera puede reconocer que exista el problema de tener que reconciliarlas; a pesar de ello, proporciona algunas de las ideas a las que ha de apelar esa reconciliación.

La moralidad contractual tiende a centrarse en la justicia y en las relaciones externas, tales como la confianza entre extraños, ¿hay algún fundamento contractual para la amistad y para las relaciones familiares genuinas? No he explorado si la moralidad de las relaciones privadas puede entenderse desde una perspectiva contractualista. Es un tema que merece atención, aunque quizá no sea mucho lo que podría añadir yo. Continuando con las relaciones externas, sí espero tener algo más que decir sobre la idea de una democracia contractual, que estoy convencido puede ofrecer un reconocimiento público de la individualidad que sirva como punto de encuentro entre una concepción de la democracia como mera coordinación de intereses privados y una concepción basada en la afirmación de valores públicos sustantivos.

Quizás el mayor foco de interés de mi trabajo reciente es la teoría de la deliberación racional. No me detendré aquí comentando los

1. «The Unity of Reason: A Subversive Reinterpretation of Kant», *Ethics*, 96 (1985), págs. 74-88. (*N. del t.*)

problemas concretos que plantean mis tesis, pero sí quiero mencionar objeciones que pueden hacerse a mi enfoque general. Decir que los procedimientos racionales de deliberación son «aquellos de los que pueda esperarse que conduzcan mejor a la realización de los propósitos del agente» supone una descripción pragmática de la racionalidad deliberativa, y tal descripción ha de afrontar muchas objeciones. Hay quienes sostienen que, así como el razonamiento sobre qué pensar o creer debe responder a un criterio objetivo de verdad, el razonamiento sobre qué hacer debe responder a un criterio objetivo de bondad. Otros, aunque aceptan la idea de que la deliberación racional depende de los propósitos del agente, insisten, sin embargo, en que los procedimientos deliberativos son racionales cuando aconsejan al agente elegir en cada ocasión la acción que conduce más directamente al logro de sus propósitos, en vez de pensar que son racionales cuando su empleo coherente promueve mejor el éxito global del agente. En «Asegurar y amenazar» respondo brevemente a ambos tipos de objeciones, pero es obvio que se trata, en el mejor de los casos, de un punto de partida.

Así pues, los ensayos recogidos en este volumen representan una tarea en marcha. Los filósofos ocultamos nuestras conjeturas bajo un manto de certeza, pero el lector no debe dejarse engañar. Estos ensayos son parte de un proyecto en marcha; así han de entenderse para que inciten al lector a responder y unirse, así, a esta conversación que llamamos filosofía.

DAVID P. GAUTHIER
Pittsburgh, diciembre de 1996

# JUSTICIA Y DOTACIÓN NATURAL: HACIA UNA CRÍTICA DEL MARCO IDEOLÓGICO DE RAWLS

## I

El problema principal de la Ética moderna es la reconciliación de la moralidad con la racionalidad. Es fácil conseguir esta reconciliación seleccionando concepciones de la racionalidad y la moralidad a la medida, de modo que se complementen. Pero tiene más valor e interés intentarlo partiendo de nuestras concepciones intuitivas de la racionalidad y la moralidad. La *Teoría de la Justicia* presume de lograrlo con respecto a la parte de la moralidad que constituye el ámbito de la justicia.

La Teoría de la Justicia de Rawls[1]

> generaliza y lleva a un nivel más elevado de abstracción la conocida Teoría del Contrato Social (pág. 28)... Según el contractualismo, la Teoría de la Justicia es parte de la Teoría de la Elección Racional (pág. 67)... La idea directriz es que los principios de justicia para la estructura básica de la sociedad son el objeto del acuerdo original. Son los principios que las personas libres y racionales interesadas en promover sus propios intereses aceptarían en una posición inicial de igualdad como definitorios de los términos fundamentales de su asociación (pág. 28).

Así pues, la conexión entre racionalidad y justicia ha de ser establecida en términos de un acto de decisión en una posición inicial de igualdad. Por tanto, además de explicar la racionalidad y la justicia, la

1. Los números entre paréntesis corresponden a la paginación de la edición española de la *Teoría de la Justicia*, México, FCE., 1979. Las citas reproducen la traducción de María Dolores González, salvo las pequeñas modificaciones necesarias para mantener la coherencia del texto. (*N. del t.*)

teoría debe ofrecer una explicación de lo que Rawls denomina normalmente la «situación inicial». La hipótesis es que estos factores pueden definirse de un modo que, por un lado, se ajuste a nuestros juicios reflexivos sobre los mismos y, por otro, los articule de manera que la concepción de la justicia quedaría incorporada en los principios elegidos por hombres racionales en la situación inicial. Es esta hipótesis la que hace que la teoría sea una verdadera *teoría*.

Rawls reconoce explícitamente la importancia de esta hipótesis en relación a la justicia y la situación inicial. Afirma que

> parece natural pensar que el concepto de justicia es distinto de las diferentes concepciones de la justicia, y que está especificado por el papel que tienen en común estos diferentes conjuntos de principios y concepciones (págs. 21-22).

La teoría debe, por tanto, ofrecer esa

> explicación del sentido de la justicia de una persona... de acuerdo con sus juicios tras un equilibrio reflexivo... Este estado se alcanza después de que una persona ha sopesado varias concepciones propuestas y, o bien ha revisado sus juicios de acuerdo a una de ellas, o bien se ha mantenido fiel a sus convicciones iniciales (y a la concepción correspondiente) (págs. 68-69).

Además,

> existen... muchas interpretaciones posibles de la situación inicial... Sin embargo... existe una interpretación de la situación inicial que expresa mejor las condiciones que se considera [razonable] imponer en la elección de los principios, condiciones que, además, conducen a una concepción que corresponde a nuestros juicios en equilibrio reflexivo. Esta interpretación más aceptada, o estándar, la llamaré la posición original (págs. 146-147).

Y así, la teoría de la

> justicia como imparcialidad es la hipótesis de que los principios que serían elegidos en la posición original son idénticos a aquellos que corresponden a nuestros juicios reflexivos y que, por tanto, esos principios describen nuestro sentido de la justicia (pág. 68).

Pero el modo en que Rawls explica la racionalidad es significativamente diferente de su explicación de la justicia y la situación inicial. No dice que le interese (la elección de) una concepción determinada de la racionalidad, entre las varias que podrían presentarse, sino el concepto mismo.

> El concepto de racionalidad invocado aquí es ... el que se usa comúnmente en la Teoría Social. Así, es usual que se piense que una persona racional tiene un conjunto coherente de preferencias entre las alternativas que se le ofrecen. Ordenará estas opciones de acuerdo con el grado en que promuevan sus propósitos; llevará a cabo el plan que satisfaga el mayor número de sus deseos y, al mismo tiempo, el que tenga más probabilidades de ejecutar con éxito (pág. 170).

Rawls añade explícitamente un supuesto especial,

> que a un individuo racional no le asalta la envidia. No es de los que están dispuestos a aceptar una pérdida para sí con tal de que otros también la sufran (pág. 170).

Esto es parte de la concepción estándar si «pérdida» se entiende en relación a los propósitos y deseos del agente. Este supuesto especial puede formularse más adecuadamente diciendo que hay ciertos *bienes primarios* y

> las personas en la posición original ... preferirían tener más [bienes primarios] que menos (pág. 114).

No hay nadie que sirva a sus propósitos aceptando menos bienes primarios para sí, aunque otros también tengan menos.

El concepto de racionalidad que se emplea normalmente en la teoría social identifica la racionalidad con la maximización de la utilidad individual. Por tanto, Rawls acepta esta identificación con la suposición especial de que hay una clase de bienes cuyo incremento siempre representa un incremento de utilidad. Esta visión de la racionalidad no es propuesta como una concepción particular, que compitiera con otras en la pretensión de ser una adecuada formulación de nuestra idea de hombre racional, sino como la expresión del concepto mismo. Por tanto, la creencia, común en el pensamiento medieval, de

que el papel de la razón es permitir que el hombre conozca la ley divina, no sólo debe ser incorrecta, sino además *conceptualmente* equivocada.

Una *ideología* puede caracterizarse, en parte, por la identificación de una determinada concepción de la racionalidad con el concepto mismo. De ahí que la aceptación rawlsiana de esa concepción de la razón dominante —y, en realidad, esencial— en nuestra sociedad, caracterice su propio *marco ideológico* y lo identifique, en un aspecto fundamental, con lo que neutralmente denominaré el marco liberal individualista. En el argumento que desarrollaré a continuación sugiero una base para la crítica de ese marco.

El argumento tiene la siguiente estructura. En la sección II presento la principal característica de la concepción rawlsiana de la justicia y su descripción de la situación inicial. En la sección III muestro cómo la concepción maximizadora de la racionalidad, en conjunción con la concepción de la posición original, parece conducir al principio de la diferencia que define la concepción rawlsiana de la justicia. Sin embargo, en las secciones IV y V muestro que una reflexión sobre la dotación natural del hombre en relación, primero, con la ausencia de sociedad y, segundo, con su potencial social, exige modificar el principio de la diferencia. En la sección VI examino el efecto de estas modificaciones sobre el tipo de sociedad instituida por la adopción de los principios de la justicia y muestro que, en contra de lo que Rawls cree, ésta resulta ser una sociedad civil, o privada. Finalmente, en la sección VII menciono una concepción alternativa de la sociedad y su relación con la dotación natural del hombre, como punto de partida para una crítica de la concepción maximizadora de la racionalidad práctica.[2]

## II

Rawls formula la concepción general de la justicia derivada de su teoría del siguiente modo:

2. La posición que aquí desarrollo como corrección de las ideas de Rawls, puede encontrarse en una formulación positiva, basada en conceptos de Teoría de Juegos, en Gauthier, «Rational Cooperation», *Nous* 8 (1974), págs. 53-65, y especialmente en Gauthier, *Moral por acuerdo*, Barcelona, Gedisa, 1994.

> Todos los bienes sociales primarios —libertad, igualdad de oportunidades, renta, riqueza, y las bases de la autoestima—, han de ser distribuidos de un modo igual, a menos que una distribución desigual de uno o de todos estos bienes redunde en beneficio de los menos aventajados (pág. 341).

De esta concepción genérica se deriva una concepción más concreta que Rawls considera aplicable a nuestras circunstancias reales. Entre los bienes primarios hay que distinguir los «derechos y libertades fundamentales» y los «beneficios económicos y sociales». Los primeros han de estar igualmente abiertos a todos; los segundos han de distribuirse de acuerdo con lo que Rawls llama el *principio de la diferencia*. En la forma más general en que Rawls lo presenta —más general incluso que la forma en que lo emplea, pero con la ventaja de ser menos restrictiva en sus suposiciones sobre las condiciones reales—, el principio dice que

> en una estructura básica con *n* representantes relevantes, ha de maximizarse primero el [bienestar] de las personas representativas de la peor situación; segundo, manteniendo un bienestar igual para estas personas, ha de maximizarse el de los que les siguen, y así sucesivamente hasta llegar a los representantes de los mejor colocados, cuyo bienestar habrá de maximizarse una vez maximizado el de los precedentes *n-1*. Éste podría considerarse el principio lexicográfico de la diferencia (pág. 105).

El principio de la diferencia, que emplearé en esta forma lexicográfica genérica, distingue lo que Rawls llama igualdad democrática de la igualdad liberal. Ambas anulan las influencias de las [contingencias sociales, pero la concepción liberal

> permitiría que la distribución de riqueza y renta fuese determinada por la distribución natural de capacidades y talentos. Las cuotas distributivas se asignan de acuerdo al resultado de una lotería natural (aunque dentro de los límites establecidos por las instituciones básicas); y [este resultado es arbitrario desde una perspectiva moral. No hay mejor razón para permitir que la distribución de renta y riqueza sea resuelta en función del patrimonio natural que por la fortuna histórica y social (págs. 95-96).

La concepción democrática, basada en el principio de la diferencia, no permite que la riqueza y la renta sean determinadas por los talentos naturales. Por tanto

> la concepción de la justicia [de Rawls] ... anula los accidentes de la dotación natural ... y expresa el resultado de rechazar aquellos aspectos del mundo social que desde un punto de vista moral parecen arbitrarios (pág. 32).

Puedo adelantar que el objetivo de mi argumento consiste en mostrar que este carácter *anulador* de la concepción rawlsiana de la justicia es incompatible con su base *contractual.* Argumentaré que si aceptamos el marco contractual implicado por la concepción maximizadora de la racionalidad que Rawls abraza, nos vemos abocados a una concepción de la justicia cercana a la concepción liberal, que Rawls, sin embargo, rechaza. En la medida en que esta concepción liberal no coincida con nuestros juicios reflexivos ni, por tanto, con nuestro sentido de la justicia, no se habrá producido la reconciliación entre la racionalidad y la moralidad.

La base contractual del pensamiento de Rawls exige, como hemos visto, una especificación de la situación inicial en la que han de elegirse los principios de la justicia. La especificación que se considere adecuada definirá la posición original. En la posición original se supone que todas las personas son racionales, por tanto se supone que tratan de maximizar su bienestar, con la restricción de que ninguna es envidiosa. Cada una está sujeta (y es consciente de ello) a lo que Rawls denomina las «circunstancias de la justicia», las cuales

> pueden describirse como las condiciones normales bajo las cuales la cooperación humana es tanto posible como necesaria... La descripción que de ellas hace Hume es particularmente lúcida... En aras de la simplicidad, subrayo a menudo la condición de moderada escasez (entre las circunstancias objetivas), y la del mutuo desinterés, o el hecho de que a los individuos no les preocupan los intereses de otros (entre las subjetivas). Así, puede decirse brevemente que las circunstancias de la justicia se dan siempre que, en condiciones de moderada escasez, personas mutuamente desinteresadas plantean demandas conflictivas sobre la división de beneficios sociales (págs. 152-154).

Las personas en la situación original tienen un buen conocimiento general sobre su situación, pero ninguno particular. Han de elegir tras un «velo de ignorancia» que asegura que

> nadie conoce su lugar en la sociedad, su posición o clase social; tampoco sabe cuál será su suerte en la distribución de talentos y capacidades naturales, su inteligencia y fuerza, etcétera. Igualmente, nadie conoce su propia concepción del bien, ni los detalles de su plan racional de vida, ni siquiera los rasgos particulares de su propia psicología, tales como su aversión al riesgo o su tendencia al optimismo o al pesimismo (págs. 163-164).

Existen algunos otros aspectos desconocidos sobre la sociedad; pero lo esencial para mi argumento puede extraerse del pasaje citado: *nadie sabe quién es.* Y como nadie sabe quién es, o en qué circunstancias estará, cada hombre es todos —o ninguno—. Así,

> es claro que, puesto que las diferencias entre las partes son desconocidas y puesto que todas son igualmente racionales y se hallan en la misma situación, todas serán susceptibles de ser convencidas por los mismos argumentos. Por tanto, podemos contemplar la elección en la situación original desde el punto de vista de una persona seleccionada al azar. Si cualquiera, después de la debida reflexión, prefiere una concepción de la justicia a otra, entonces todos la preferirán, y se alcanzará un acuerdo unánime (pág. 166).

Esto describe suficientemente la posición original. La tarea siguiente es mostrar por qué se supone que personas racionales en esta posición elegirían la concepción de la justicia de Rawls y, más en concreto, el principio lexicográfico de la diferencia.

## III

En este examen de la derivación del principio de la diferencia como objeto de la elección de personas racionales bajo las circunstancias de la justicia en la posición original, no he seguido exactamente el argumento de Rawls. El análisis de ese argumento requeriría un ensayo entero y revelaría, creo, ciertas debilidades que mi

propia descripción ha tratado de evitar. Así que si me aparto de Rawls no es para desmerecer su tesis, y demostrar más fácilmente que es vulnerable a la crítica, sino todo lo contrario; para presentar el principio lexicográfico de la diferencia bajo la perspectiva más favorable posible.

Es esencial observar la derivación del principio desde dos puntos de vista diferentes. Primero, por supuesto, está el punto de vista de quien razona en la posición original y elige «tras un velo de ignorancia» (pág. 29). En segundo lugar está el punto de vista de quien razona después que el principio es elegido, la sociedad resultante instituida, y el velo de ignorancia levantado. Mi opinión es que un principio que establece los términos de una asociación debe ser racional no sólo prospectivamente, sino retrospectivamente. Con esto no quiero decir que la elección hecha en condiciones de ignorancia sea la que uno haría, si pudiera, cuando ya es consciente de las circunstancias reales que determinan cómo le va finalmente en la sociedad, sino que el conocimiento de ese resultado no debería hacernos considerar irracional la decisión original, cuando nos situáramos heurísticamente en las condiciones bajo las que fue tomada. Aquella decisión debe ser racionalmente aceptable, en este sentido, para todas las personas representativas una vez que se levanta el velo de ignorancia. Me parece que este requisito está implícito en la tesis rawlsiana de que «la posición original se interprete de tal modo que uno pueda adoptar su perspectiva en cualquier momento» (pág. 166).

Consideremos primero el método de decisión racional de una persona que sabe quién es. Tal persona elegiría los principios que maximizaran su beneficio esperado. Asignaría a cada conjunto posible de principios un futuro esperado, que sería la media ponderada de los posibles futuros que podrían resultar de la adopción de esos principios. Después asignaría a cada futuro esperado un valor o utilidad, que indicaría el beneficio esperado de ese futuro para ella. Y elegiría entonces los principios que maximizaran este valor.

Pero en la posición original nadie sabe quién es, así que lo que tenemos que intentar formular es el razonamiento de todos y cada uno. Si todos razonaran de forma análoga a como lo hace una persona particular, asignarían a cada conjunto posible de principios un conjunto de futuros esperados —cada miembro del conjunto representaría el futuro esperado para cada persona representativa, cada una de la

personas que cualquiera podría resultar ser—. A cada uno de estos futuros esperados le asignarían un valor, que reflejaría el beneficio esperado del mismo para la persona en cuestión. Calcularían después el valor esperado de cada conjunto de futuros esperados: la media ponderada de los valores de sus miembros (aquí la ponderación correspondería a la probabilidad de ser cada una de las personas posibles). Este último valor esperado es lo que maximizarían al elegir los principios. Y según esto, elegirían el principio de la utilidad media, y no el principio de la diferencia.

Rawls discute ese tipo de utilitarismo con cierta amplitud (sección 28), pero sin plantear algunas objeciones al mismo que, desde mi punto de vista, son absolutamente convincentes. Me parece que el mejor argumento contra él, y en favor del principio de la diferencia, proviene de considerar la situación de la persona que, al levantar el velo de ignorancia, encuentra que es la menos aventajada. Salvo que todos estén igual, siempre habrá una persona representativa que encuentre, por un lado, que nadie está peor que ella y, por otro, que su utilidad es menor que la media ponderada que esperaba ¿Cómo va a autoconvencerse de que su elección de principios fue racional?

Mientras la mala suerte dependa de circunstancias de hecho, no se plantea ningún problema real. Se podía esperar que las circunstancias hubieran sido diferentes, más favorables a su bienestar. La persona desafortunada puede imaginarse a sí misma en esas otras circunstancias, disfrutando de mayor bienestar. Esa otra alternativa esperada hace que pueda aceptar su sacrificio real en términos de bienestar: no hubo suerte, eso es todo; pero hay que admitir que, dada la información de que disponía, la elección fue la mejor posible.

Sin embargo, en la medida en que la mala fortuna dependa de la identidad de la persona, la situación en que se encuentra es muy diferente. Porque, incluso aunque sea inteligible la suposición de que pudiera haber sido otra persona con más suerte, no puede imaginarse a *ella misma* siendo esa otra persona que disfruta de mayor bienestar. Claro que puede imaginarse a ella misma en las circunstancias de otra persona, pero ésta no es la alternativa que esperaba, porque *ex hipothesi* su bienestar y, por tanto, sus circunstancias van a depender de quién sea. También puede imaginar a alguna otra persona disfrutando de mayor bienestar —de hecho ésta es la alternativa esperada relevante—, pero esto no hace más aceptable su sacrificio presente en

términos de bienestar. Ahora que sabe quién es, ya no supone que la elección que hizo fue la mejor para ella. Al contrario, se da cuenta de que la ignorancia sobre quién era la incapacitó para realizar la mejor elección posible para sí misma.

Como he dicho, este argumento no se encuentra en Rawls. Aunque sí lo sugiere en parte (págs. 202-204), especialmente cuando sostiene que

> la expectativa sobre la cual descansa el razonamiento en pro del principio de la utilidad media es defectuosa... supone que las partes no tienen ni un carácter ni una voluntad definidas, no son personas con un sistema de fines, sobre el que debe estar fundada toda expectativa (pág. 175).

Si todos eligen los principios que conducen a la maximización de la utilidad media esperada, *no se sigue* que estén eligiendo aquellos principios que maximizarían la utilidad esperada de cualquiera de las personas representativas que uno puede resultar ser. Evidentemente, nadie podría elegir principios que maximizaran la utilidad esperada de todas esas personas, porque el resultado que maximizara la utilidad esperada de una persona representativa no tiene por qué coincidir, en general, con el que maximizara la de cualquier otra. ¿Qué alternativa queda abierta, entonces, para *todos*?

Todos deben elegir unos principios tales que cuando se levante el velo de ignorancia, cada persona representativa considere que ella habría elegido racionalmente esos mismos principios. Dada la conexión entre racionalidad y maximización del beneficio que venimos suponiendo, una condición que la elección de principios debe satisfacer es que el resultado determinado por ellos sea tal que ninguna persona representativa pueda esperar recibir mayores beneficios sin que otra persona tenga que renunciar a parte de los suyos. Porque si alguien encuentra que podría haber esperado estar mejor sin coste para nadie, entonces no consideraría que ha elegido racionalmente. Así pues, la optimidad en el sentido de Pareto, o la eficiencia, como Rawls prefiere denominarla, es una condición de la decisión de todos.

Pero normalmente hay muchos resultados óptimos y, por tanto, muchos principios que satisfarían la condición. Se necesita, en consecuencia, una segunda condición, y ésta surge de la igualdad inicial de

todas las personas. Dado que la elección ha de ser aceptable para cada persona representativa, no se pueden elegir principios que favorezcan el bienestar de unas personas a costa de otras. Aunque las personas no sean envidiosas y, por tanto, no valoren los beneficios que reciben comparándolos con los que reciben otras, la aceptabilidad racional de un principio que asigne a una persona concreta un conjunto dado de beneficios esperados, puede depender de la relación entre ese conjunto y los conjuntos de beneficios asignados a otras personas. Una persona no considerará que su decisión es racionalmente aceptable si los principios elegidos le asignan beneficios esperados menores que los de otra persona, excepto en la medida en que la desigualdad no pudiera ser rectificada incrementando sus propios beneficios sino *únicamente* reduciendo los beneficios de la persona más favorecida.

Por tanto, la persona menos aventajada considerará que la elección de principios es racional sólo si no existen principios alternativos que pudieran proporcionarle mayores beneficios (aunque siguiera siendo la persona menos aventajada), ni modo alguno de dejar de ser la persona menos aventajada sin que otro viniera a quedar en una posición tan desventajosa como la que ella ahora ocupa. En consecuencia, todos deben elegir los principios que aseguren que a la persona menos aventajada le va lo mejor posible. Después habría que considerar la posición del segundo menos aventajado, y así sucesivamente. Parece evidente que todos deben elegir el principio lexicográfico de la diferencia, como procedimiento para determinar la distribución de los beneficios entre los miembros de la sociedad. En otras palabras, todos deben seleccionar aquellos principios cuya adopción instituye una sociedad en la que se producen los beneficios óptimos, los cuales se distribuyen igualitariamente si la distribución igual es óptima y, si no lo es, se distribuyen de tal modo que la persona representativa menos favorecida obtenga el mejor resultado posible, compatible con que la segunda menos favorecida obtenga el mejor resultado posible, y así sucesivamente. De este modo parece sostenerse la teoría rawlsiana sobre la derivación de su concepción de la justicia —que incluye el principio de la diferencia— a partir de la concepción maximizadora de la racionalidad y una posición original de ignorancia. No obstante, en las próximas dos secciones sugiero que Rawls no describe satisfactoriamente cómo habría de aplicarse el

principio lexicográfico de la diferencia a la distribución de beneficios, y que una explicación más satisfactoria nos hace ver el principio bajo una luz bien distinta.

## IV

El principio lexicográfico de la diferencia va a regular la distribución de los bienes sociales primarios clasificados como beneficios económicos y sociales. Algunos de estos bienes son una creación de la sociedad. Pero otros, en concreto los beneficios económicos, deben ser accesibles, en alguna medida, bajo cualquier circunstancia. Podríamos preguntar, por tanto, cómo se distribuirían estos bienes en ausencia de acuerdo sobre un principio regulador de su asignación. ¿Qué pasaría si no hay contrato social?

Preguntar esto equivale a preguntar por el «punto de no acuerdo» (pág. 174). Rawls distingue dos formas de interpretar este punto, a las que denomina respectivamente «egoísmo general» y «estado de naturaleza». Supone que la posición original corresponde al estado de naturaleza (pág. 29), pero mantiene que el «egoísmo general» sería el resultado en caso de no haber acuerdo. Desafortunadamente, no describe el egoísmo general en términos del comportamiento de personas racionales. No obstante, supondré que egoísmo general se refiere a la situación en que cada persona procura simplemente asegurarse tanto como sea posible para sí misma: la distribución de los bienes sociales primarios, en la medida en que existan aparte del acuerdo y de la sociedad por él instituida, sería el resultado de la actividad egoísta de todas las personas.

No se puede esperar que esta distribución fuera igualitaria. Las personas que forman la sociedad difieren en su dotación natural, y las diferencias en dotación natural afectarían diferencialmente la capacidad de la gente para procurarse bienes económicos en una situación de egoísmo generalizado. Por tanto, el «punto de no acuerdo» proporcionaría diferentes niveles de bienestar a las distintas personas, si bien el nivel sería bajo para todos.

Tras el velo de ignorancia nadie conoce sus capacidades y talentos naturales y, por tanto, nadie sabe qué obtendría en ausencia de acuerdo. Sin embargo, todos saben que tienen ciertas capacidades y talen-

tos naturales, y que la gente difiere en esta dotación, de tal modo que en ausencia de acuerdo cada uno podría alcanzar un nivel diferente de bienestar. Así, todos pueden tener en cuenta en su razonamiento el «punto de no acuerdo», aunque ninguna persona sepa cómo le afectaría a ella en concreto.

Las circunstancias de la justicia proporcionan la razón del acuerdo. La gente «ve la sociedad como una empresa cooperativa para el beneficio mutuo» (pág. 555): mediante la cooperación se pueden producir más bienes primarios, que todos desean. Voy a denominar a estos bienes primarios producidos mediante la cooperación, el *excedente social*; se trata de la porción de la cantidad total de bienes sociales primarios que no se producirían sin cooperación. El contenido del acuerdo debe garantizar la producción de este excedente social y determinar su modo de distribución.

Se pueden distinguir tres modos de determinar la distribución posible del excedente social:

> 1. La distribución puede compensar el desequilibrio natural de bienes primarios que cada uno se aseguraría en virtud de sus talentos y habilidades naturales en condiciones de egoísmo general. La cantidad total de bienes primarios se distribuiría de modo que la participación menor fuese máxima, de acuerdo al principio lexicográfico de la diferencia.
>
> 2. La distribución puede ignorar el desequilibrio natural de bienes primarios y maximizar la participación menor en el excedente social. El principio lexicográfico de la diferencia seguiría regulando la distribución, pero no se aplicaría a la cantidad total de bienes sociales primarios, sino sólo a la cantidad producida como resultado de la cooperación social.
>
> 3. La distribución puede ser proporcional a las participaciones de bienes sociales primarios que cada uno lograría por su cuenta en condiciones de egoísmo general.

La tercera posibilidad constituye, a mi juicio, la interpretación más plausible de la aristocracia natural —si bien esta interpretación difiere de la de Rawls (pág. 96)—. Desde este punto de vista, la sociedad debe beneficiar a cada persona de acuerdo a sus talentos y capacidades naturales, cuyo valor es estimado según el éxito que cada quien tenga en la tarea de mantenerse por su cuenta en ausencia de socie-

dad. En vez de anular los accidentes de la dotación natural, este modo de distribución hace de esos accidentes la base del mérito social.

La primera posibilidad es lo contrario de esta tercera; en vez de tomar la dotación natural como base del mérito, la considera inmerecida. Así, exige que las desigualdades naturales sean reparadas, en la medida en que pueda lograrse sin renunciar a una distribución óptima. Y éste es el modo de distribución que Rawls defiende:

> Podemos observar que el principio de la diferencia concede cierta importancia a las consideraciones determinadas por el principio de la compensación. Este principio afirma que las desigualdades inmerecidas requieren una compensación; y dado que las desigualdades de nacimiento y dotación natural son inmerecidas, habrán de ser compensadas de algún modo... Ahora bien, que yo sepa, nadie ha propuesto el principio de la compensación como el único criterio de la justicia... Pero, sean cuales fueren los otros principios que mantenemos, deberán tomarse en cuenta las demandas de la compensación. Se piensa que representan uno de los elementos de nuestra concepción de la justicia... Así pues, aunque el principio de la diferencia no es igual al de la compensación, sí realiza algunos de los objetivos de éste... El principio de la diferencia representa, en efecto, un acuerdo en el sentido de considerar la distribución de talentos naturales, en ciertos aspectos, como un acervo común, y compartir los beneficios de esta distribución, cualquiera que sea de hecho (págs. 123-124).

Sin embargo, sostener que las desigualdades naturales son inmerecidas es evidentemente erróneo. Desde luego, no son merecidas, no corresponden a mérito alguno; pero por lo mismo tampoco son inmerecidas, no son contrarias al mérito. Las desigualdades naturales —las que se ponen de manifiesto en la distinta habilidad de la gente para lograr beneficios por su cuenta en la posición de no acuerdo— deberían ser ignoradas al distribuir el excedente social.

La compensación o anulación de los accidentes de la dotación natural, no forma parte de ningún principio que pudieran acordar personas interesadas en avanzar sus propios intereses para regular la distribución del fruto de su cooperación. Como condición para aceptar libremente cooperar con los demás en términos de igualdad, cada persona esperará una participación igual en los beneficios deri-

vados de la cooperación. Cuando el velo de ignorancia se levanta y cada uno es consciente de qué bienes sociales primarios podría haber esperado en situación de egoísmo general y qué bienes puede esperar como resultado de la sociedad instituida por el acuerdo, entonces cada uno deseará que le correspondan tantos bienes *adicionales* como sea posible. Una persona no considerará racional su asentimiento si su participación en esos bienes adicionales es menor que la de cualquier otra persona, a no ser (puesto que nadie es envidioso) que una mayor igualdad sólo pueda conseguirse reduciendo la cuota de esa otra persona, y no incrementando la propia. Así, la persona que reciba la parte más pequeña del excedente social, sólo considerará el acuerdo racionalmente aceptable si no es posible que la persona menos favorecida reciba una participación mayor. Por tanto, el principio acordado debe exigir la maximización de la participación más pequeña en el excedente social, y la maximización de la segunda menor que sea compatible con lo anterior, y así sucesivamente; debe ser, en definitiva, el principio lexicográfico de la diferencia aplicado estrictamente al excedente social, sin consideración alguna a las desigualdades en la distribución de los beneficios que se darían en ausencia de acuerdo.

Un individuo maximizador-de-utilidad no considerará racional aceptar un nivel inferior de bienestar en la sociedad sólo para maximizar el beneficio promedio; ésta es una de las objeciones decisivas contra el utilitarismo. Pero tampoco considerará racional aceptar una participación menor en los beneficios de la cooperación —el excedente social— sólo para incrementar la igualdad general. La concepción de la racionalidad que Rawls acepta acríticamente llevaría a acordar principios de cooperación que maximizaran el mínimo *social*.

## V

Entre las concepciones de la justicia criticadas por Rawls se encuentra la idea de

> que la justicia significa que la sociedad tiene que hacer la misma contribución proporcional a la realización, por parte de cada persona, de la mejor vida de la que ésta es capaz (pág. 563).

Rawls cree que esta concepción, que encuentra en los escritos de Frankena y Findlay (págs. 563 y sigs.), es inaceptable no sólo por las dificultades prácticas para aplicar el criterio, sino por otra razón:

> Una dificultad más importante consiste en que las mayores facultades de algunos puedan darles un derecho mejor sobre los recursos sociales, sin compensar a otros por esas ventajas. Debe suponerse que las variaciones en las facultades naturales influirán en lo que es necesario para facilitar una asistencia proporcional igual a quienes tienen diferentes proyectos de vida. Pero, además de violar el principio del beneficio mutuo, esta concepción de la igualdad implica que la solidez de las demandas de cada uno depende directamente de la distribución de las capacidades naturales, y por tanto de contingencias que son arbitrarias desde un punto de vista moral (págs. 563-564).

Si suponemos que nuestro sentido de la justicia está representado por los principios que acordarían maximizadores-de-utilidad racionales, entonces hemos de admitir que la violación del principio del beneficio mutuo es inaceptable. De hecho, la idea de beneficio mutuo constituye uno de los argumentos que respaldan nuestra tesis de que el principio lexicográfico de la diferencia ha de aplicarse al excedente social; si se aplicara a la suma total de los bienes primarios, sería en principio posible que alguna persona muy favorecida en condiciones de egoísmo general no recibiera ningún beneficio de la sociedad. Pero por aplicar el principio de la diferencia a la contribución proporcional hecha por los bienes sociales primarios al bienestar de cada persona —en vez de a la cantidad absoluta de ellos que cada persona representativa recibe—, no se viola el principio del beneficio mutuo. La cuestión es cómo medir la participación de una persona en los bienes sociales primarios: ¿debe el principio distributivo maximizar la *cantidad* mínima de esos bienes, o debe maximizar la *proporción* mínima en que se realiza la mejor vida posible de la persona?

Las diferencias en dotación natural influirán no sólo en el nivel de beneficios que la gente puede alcanzar por su cuenta en ausencia de acuerdo, sino también en los niveles posibles de beneficio que pueden lograrse mediante la cooperación. Así pues, vamos a intentar desarrollar una medida del efecto social potencial de las diferencias en dotación natural.

Consideremos en abstracto todos los posibles sistemas de producir y distribuir un excedente social. Las distribuciones negativas no estarían permitidas; toda persona debe esperar, bajo cualquier sistema, beneficios que, al menos, igualen los que esperaría obtener en ausencia de acuerdo. Con esto ya se satisface un requisito mínimo de beneficio mutuo. Ahora bien, cabe suponer que [los distintos sistemas afectarán tanto a la cantidad total de excedente producido como a la cuota del mismo que recibe cada persona representativa] Podemos restringir nuestra atención a los sistemas óptimos, o eficientes; y concluir que para cada persona representativa habrá un sistema óptimo que maximize su bienestar.

Las diferencias en talentos y capacidades naturales se manifestarán en forma de diferencias en el nivel máximo de bienestar que pueden alcanzar las distintas personas. Si comparamos el bienestar que corresponde a la persona naturalmente inteligente, fuerte y emprendedora, en el sistema que más favorezca a este tipo de personas, con el bienestar que correspondería a las personas naturalmente torpes, débiles y perezosas, en el sistema que más les favorezca, encontraremos que el primero será mayor. De hecho, quizás encontremos que a la persona naturalmente bien dotada le va mejor que a la menos dotada, incluso en el sistema que maximiza los beneficios de éstos últimos, porque muy bien puede suceder que para maximizar la cantidad absoluta de bienes sociales correspondientes a los menos capaces, haya que recompensar con una cantidad mayor de estos bienes a los mejor dotados, que son los mayores responsables de la producción del excedente social.

Al máximo bienestar que cada persona representativa puede esperar bajo el sistema social óptimo más favorable para ella, lo denominaré su *potencial social*. El potencial social es, por tanto, una medida de los efectos sociales potenciales de las diferencias en dotación natural. Tras el velo de ignorancia nadie conoce cuál es su potencial social, pero sí se sabe que todos tienen un potencial social y que éste difiere de persona a persona. De ahí que dicho potencial pueda ser tenido en cuenta por todos en su razonamiento.

Se pueden distinguir tres modos de determinar la distribución posible del excedente social, teniendo en cuenta el potencial social:

1. La distribución puede ignorar el potencial social, de modo que el excedente social se distribuya directamente de acuerdo con el prin-

cipio lexicográfico de la diferencia. La interpretación rawlsiana del principio de la diferencia se apartaría de ésta sólo en que él habla de distribuir todos los bienes sociales primarios, y no sólo el excedente social.

2. Se puede distribuir el excedente social de modo que se maximize el grado mínimo en el cual se realiza el potencial social de cualquier persona; así el principio lexicográfico de la diferencia se aplicaría a la contribución relativa de los bienes sociales primarios al bienestar de cada uno.

3. Se puede tomar el potencial social como medida del mérito, de modo que la distribución del excedente social maximizara la satisfacción de la persona con mayor potencial social, y así sucesivamente, en orden lexicográfico descendente.

De nuevo, la tercera de estas posibilidades parece avalar una forma de aristocracia; podríamos llamarla «aristocracia social», para distinguirla de la aristocracia natural mencionada en la sección anterior. Los talentos y capacidades naturales se toman como indicativos del mérito, y son recompensados. Este punto de vista se opone a nuestra concepción intuitiva de la justicia y, además, no sería aceptado por personas racionales en condiciones de imparcialidad.

Pero tales personas considerarían la primera posibilidad igualmente insatisfactoria, porque tiende a anular los efectos de las capacidades naturales, tratándolos como inmerecidos. Debe recordarse que estamos considerando qué principios de asociación serían adoptados por personas libres y racionales, preocupadas por promover sus intereses particulares. Cada una espera recibir tanto beneficio de la sociedad como cualquier otra (excepto cuando el único modo de lograr esa igualdad sea la reducción de los beneficios de otros, sin aumento de los propios). Pero ¿cómo medirá cada uno el beneficio que recibe? Mi tesis es que el potencial social proporciona la medida natural. Si una persona casi logra la mejor vida de la que es capaz, y otra recibe una cantidad de bienes sociales que no le permiten alcanzar más que una pequeña fracción de su mejor vida potencial, juzgaríamos que la primera persona se beneficia más, cualquiera que sea la cantidad efectiva de bienes sociales primarios recibidos por cada cual.

Pensemos una vez más en lo que ocurre al levantar el velo de ignorancia. Según la interpretación rawlsiana del principio de la diferencia, las dotaciones naturales han de tomarse como acervo común,

de modo que la relación entre beneficio recibido y talento y esfuerzo empleados, descenderá conforme el talento y el esfuerzo se incrementan. De ahí que el hombre naturalmente favorecido percibirá que sus talentos y esfuerzos se dirigen principalmente a beneficiar a los naturalmente menos capaces. Pero no descubrirá ninguna razón para no tener derecho a beneficiarse de sus propias capacidades en la misma medida que los demás. Cualquier sistema que no le garantice este derecho será percibido como negativamente discriminatorio en relación a sus capacidades naturales. Así, no considerará racional acordar principios que le exijan aceptar un beneficio proporcional menor simplemente para que el beneficio absoluto sea igual en la sociedad.

Mi conclusión es, por tanto, que personas racionales en el sentido aceptado por Rawls, sujetas a las condiciones de la justicia en una posición de igualdad original, acordarán un sistema de cooperación social regulado por un principio de la diferencia que especifique que cada uno recibirá tanto beneficio como esperaría al margen del acuerdo y, además, que se asignará a cada uno la participación en el excedente social de bienes primarios que asegure la maximización del nivel mínimo de realización del potencial social y, garantizados niveles mínimos iguales, la maximización del segundo menor nivel de realización del potencial social, y así sucesivamente. El objeto del acuerdo racional es el principio lexicográfico de la diferencia aplicado a una distribución del excedente social proporcional al potencial social. En adelante me referiré a éste como el *principio de la diferencia proporcional*.

## VI

El principio de la diferencia proporcional da lugar a un tipo de sociedad muy diferente del previsto por Rawls. Rawls supone que «una sociedad está bien ordenada no sólo cuando está diseñada para promover el bien de sus miembros, sino cuando, además, está efectivamente regulada por una concepción pública de la justicia» (pág. 21). Rawls arguye que los principios elegidos en la posición original dan lugar a una sociedad bien ordenada porque, además de ser aceptados por todos los miembros de la sociedad como base de su conducta, están incorporados en las instituciones de la sociedad. Esto le

permite afirmar que la justicia como imparcialidad es coherente con esa idea de sociedad (bien ordenada). Las enmiendas que he hecho al principio de la diferencia no afectan esta descripción. El principio de la diferencia proporcional como objeto de acuerdo racional en la posición original, puede igualmente dar lugar a una sociedad bien ordenada en la que el bien de sus miembros se garantiza mediante la producción óptima de beneficios, y se cumple con la justicia al distribuir esos beneficios de un modo que maximiza el nivel mínimo de satisfacción. La concepción pública de la justicia difiere de la defendida por Rawls, pero no hay razón para suponer que sería menos eficaz en la regulación de las actividades de los individuos y la estructura de las instituciones de la sociedad.

Pero Rawls supone además que

> los principios de la justicia se relacionan con la sociabilidad humana. La idea principal consiste, sencillamente, en que una sociedad bien ordenada (correspondiente a la justicia como imparcialidad) es, en sí misma, una forma de unión social. En realidad, es una unión social de uniones sociales. Concurren en ella dos rasgos característicos: el positivo funcionamiento de instituciones justas es el fin último compartido por todos los miembros de la sociedad, y estas formas institucionales son apreciadas como buenas en sí mismas (pág. 582).

Lo que permite a Rawls distinguir el tipo de sociedad instituida mediante la elección de principios que incorporan una concepción de justicia como imparcialidad, de la que denomina «sociedad privada».

> Sus principales rasgos son: primero, que las personas que la forman... tienen sus propios fines privados, que son contrarios o independientes, pero en ningún caso complementarios; y, segundo, no se considera que las instituciones tengan valor alguno por sí mismas... Así, cada persona valora la organización social sólo como un medio para sus objetivos privados... Cada cual prefiere el esquema más eficaz, que le dé la mayor proporción de beneficios (pág. 576).

A lo que añade que «la teoría de los mercados competitivos es una descripción paradigmática de este tipo de sociedad» (pág. 576).

La sociedad privada, o civil, como Hegel la denomina (según apunta el propio Rawls), es sin duda la forma de sociedad descrita en

las teorías contractualistas clásicas de Hobbes y Locke. Sin embargo, Rawls quiere disociar su uso del modelo contractualista de su conexión histórica con este tipo de sociedad. Después de todo, el paradigma de Rawls es Kant, no Hobbes. En ese sentido, afirma lo siguiente:

> Se sostiene, a veces, que la doctrina contractualista implica que la sociedad privada es el ideal, por lo menos cuando la división de beneficios satisface una adecuada norma de reciprocidad. Pero esto no es así, como la noción de una sociedad bien ordenada demuestra (pág. 577).

Sin embargo, este argumento es falaz. Lo que debe oponerse a la noción de una sociedad privada no es la concepción de una sociedad bien ordenada, sino una sociedad basada en la justicia como imparcialidad. Es obvio que una sociedad bien ordenada no es necesariamente una sociedad privada pero, como mostraré enseguida, una sociedad privada puede ser una sociedad bien ordenada. El argumento de Rawls debería ser, según esto, que la doctrina contractualista no implica una sociedad privada porque en un estado de naturaleza adecuadamente definido, el producto del contrato es una sociedad regulada por la concepción de la justicia como imparcialidad. Pero el argumento expuesto en las secciones anteriores de este artículo ha mostrado que el contrato social no conduce a la concepción de la justicia expresada en el principio rawlsiano de la diferencia, sino más bien a la concepción expresada por el principio de la diferencia proporcional. Por tanto, debemos preguntar si una sociedad basada en este principio es una unión social o por el contrario una sociedad privada.

La respuesta es clara. El principio de la diferencia proporcional describe exactamente una sociedad de mercado competitivo, el paradigma de la sociedad privada. Porque el mercado competitivo es el mecanismo mediante el cual se produce un excedente social óptimo, y es distribuido de acuerdo a la contribución hecha por cada persona. La única restricción que impone el principio de la diferencia proporcional es que la posición inicial de cada persona en el mercado debe reflejar sólo su dotación natural, y no factores tales como la riqueza heredada. Lo cual equivale a decir que el principio lleva el individualismo hasta sus últimas consecuencias, pues niega incluso la

relación de herencia que, aunque se mantiene en las sociedades de mercado existentes, contradice de hecho los supuestos individualistas que constituyen el fundamento de estas sociedades.

Si se acepta que, efectivamente, el principio de la diferencia proporcional es el que sería acordado entre personas libres y racionales en una posición inicial de igualdad, entonces la doctrina del contrato, bien interpretada, sí exige una sociedad privada. Además, dado que el principio de la diferencia proporcional conduce a una sociedad bien ordenada, la doctrina contractualista exigiría una sociedad privada bien ordenada. No todas las formas de sociedad privada son necesariamente bien ordenadas; una sociedad de mercado en la que la posición inicial de cada persona dependa ampliamente de su herencia podría no satisfacer una concepción pública de la justicia. Y obviamente no todas las sociedades bien ordenadas son privadas. Pero la unión social, o cualquier otra sociedad bien ordenada no privada, no puede derivarse de un acuerdo entre personas racionales en las circunstancias de la justicia.

En este punto cabría plantear una objeción al presente argumento. Un defensor de Rawls que aceptara que el principio de la diferencia proporcional se deriva de las premisas rawlsianas, podría sostener que la consecuencia de esto es que hay que modificar la descripción de la situación inicial formulada por Rawls. El principio de la diferencia proporcional se deriva suponiendo que en la situación inicial todos saben que hay diferencias de dotación natural aunque nadie sabe cuál es su propia dotación. Lo que habría que hacer para llegar a la conclusión de Rawls, y compatibilizar así la idea de unión social con la doctrina contractualista, es rediseñar la situación inicial de modo que las partes no dispusieran de este conocimiento.

Pero ¿sería satisfactoria tal modificación de la descripción de la situación inicial? La posición original ha de asegurar que la elección que gente racional realizaría en su contexto continuará siendo racionalmente aceptable cuando se levante el velo de ignorancia. Es evidente que las personas racionales, preocupadas por promover sus intereses particulares, no considerarán que un conjunto de instituciones es aceptable a menos que éstas reconozcan las diferencias en dotación natural. Dado que la sociedad se ve como un medio para producir y distribuir un excedente social óptimo, las personas racionales sólo aceptarán principios distributivos que se circunscriban al excedente, y

que lo distribuyan en función de la contribución de cada uno a su producción.

Lo que obstaculiza la derivación de la concepción rawlsiana de la justicia no es, por tanto, la descripción de la situación inicial, sino más bien su concepción de la racionalidad. El principio de la diferencia proporcional es el resultado necesario de la concepción maximizadora de la racionalidad; y es, a la vez, la base de la sociedad privada, una forma profundamente individualista de sociedad de mercado competitiva. La concepción instrumental de la racionalidad, que Rawls adopta, conduce a una concepción instrumental de la sociedad humana, que él desea rechazar. El marco ideológico de Rawls conduce a una concepción individualista liberal de la justicia y la sociedad, no a la igualdad democrática ni a la idea de unión social.

## VII

El principio de la diferencia proporcional expresa la idea de justicia derivada de la concepción maximizadora de la racionalidad, esencial en nuestra sociedad. Si la racionalidad y la moralidad han de reconciliarse dentro de nuestro marco ideológico y social, ésta es la forma que debe adoptar esa reconciliación, en lo que respecta al ámbito de la justicia. No es la reconciliación que Rawls propone, pero esto no prueba nada en contra de ella. Si conceder a cada persona los beneficios que maximicen el nivel mínimo de satisfacción conduce a conclusiones que concuerden con nuestros juicios en equilibrio reflexivo, ello bastaría para tener una alternativa adecuada a la teoría rawlsiana de la justicia.

Sin embargo, no he argumentado, ni pienso hacerlo en este momento, a favor o en contra de la concepción revisada de la justicia incorporada en el principio de la diferencia proporcional. Ahora bien, supongamos que la encontráramos inadecuada. Supongamos que encontráramos que la justicia como imparcialidad se adecua mejor a nuestros juicios reflexivos que la otra opción. A pesar de ello, la justicia como imparcialidad resulta incompatible con la concepción maximizadora de la racionalidad ¿Qué alternativa nos queda?

Creo que no hay sino tres. La primera es mantener la concepción de la justicia como imparcialidad mientras se admite que es no sólo

independiente de la racionalidad, sino incompatible con ella —irracional—. Esta opción es, quizás, heroica; pero absurda. La segunda es revisar nuestra concepción de la justicia, y por tanto de la moralidad, para acomodarla a la concepción maximizadora de la racionalidad. Se abandona el intento de lograr el equilibrio reflexivo, y simplemente se aceptan los juicios sobre la justicia derivados del principio de la diferencia proporcional, por mucho que se aparten de nuestras intuiciones. Ésta es la única respuesta posible dentro del marco ideológico implícitamente aceptado por Rawls y por otros defensores de teorías económicas y sociales bien conocidas. La tercera posibilidad es reconocer que la visión maximizadora de la racionalidad no es, tal como Rawls la describe, el concepto de racionalidad, sino sólo una concepción más; y tratar de incorporar nuestra concepción de racionalidad al equilibrio reflexivo en que desarrollamos nuestra concepción de moralidad. Pero esta alternativa implica trascender los horizontes conceptuales de nuestra sociedad. Este ejercicio cuestionaría lo que, no sólo la teoría social, sino también la práctica social acepta como el concepto común de racionalidad y, por ello, cuestionaría nuestra conciencia cotidiana de nosotros mismos como individuos sociales.

Esta conciencia cotidiana supone que el individuo, con su dotación natural, se concibe como esencialmente distinto de la sociedad, por muy dependiente que pueda ser de ella para su existencia y bienestar de hecho. Supone que la sociedad se concibe instrumentalmente, en tanto que sociedad privada, como un medio para la satisfacción de los individuos caracterizados asocialmente. Supone que la actividad humana ha de entenderse esencialmente en términos del logro del bienestar individual. Y, en consecuencia, la racionalidad práctica se concibe —y debe concebirse— como la actividad consistente en determinar qué hacer para maximizar las ganancias individuales. En una palabra, el estatuto asignado al individuo y su dotación natural, la concepción de la sociedad y la concepción de la racionalidad, se ensamblan en una concepción unificada de lo que es ser humano, basada en el marco ideológico individualista liberal.

Quizás el punto más débil en esta concepción es la escisión entre la dotación natural de los hombres y la estructura de la sociedad —una escisión que juega un papel tan importante en la derivación del principio de la diferencia proporcional—. Sin negar que hay diferen-

cias extrasociales entre las personas, basadas en la estructura genética, podemos suponer que tienen mucha menor importancia en la caracterización de un individuo que las diferencias socialmente determinadas. En vez de pensar en las capacidades y talentos de un individuo como una dotación natural que trae a la sociedad, se puede pensar en ellas como una creación social en su mayor parte. El individuo no se concibe, desde esta perspectiva, aparte de la sociedad, sino más bien como producto suyo. Y la sociedad no se concibe fundamentalmente como un instrumento para el bienestar de los individuos, sino como el marco dentro del cual se crean los seres humanos con todas sus diferencias características en cualidades mentales, físicas y emocionales. La actividad humana se entiende así, principalmente, como la actividad creadora de seres humanos —no simplemente mediante la reproducción física, como es obvio, sino mediante el desarrollo social de la individualidad humana—. La forma más elevada de esta actividad es la autodeterminación, la formación propia a través de la constitución de la sociedad. Y la racionalidad práctica puede entonces concebirse, no como el proceso de maximizar las ganancias individuales, sino como la permanente actividad de autodeterminación consciente.

Hay, por supuesto, otros modos de concebir la racionalidad práctica. No me interesa tanto proponer una alternativa concreta a la noción de Rawls, cuanto sugerir que la cuestión de nuestra concepción de la racionalidad debe ponerse en la agenda de la filosofía moral. La argumentación de Rawls, con su oportuno rigor, en comparación con los argumentos de la mayoría de los filósofos morales, nos permite ver que el intento de reconciliar moralidad y racionalidad conduce a una crítica de la racionalidad práctica. Y tal crítica no puede obviar un examen del marco ideológico en el que encontramos nuestras concepciones de hombre y sociedad, razón y moralidad.

# EL EGOÍSTA INCOMPLETO

## I. *¿Qué puede hacer un egoísta?*

1. «El egoísmo ... es la doctrina que sostiene que cada uno de nosotros debe perseguir como fin último la mayor felicidad posible para sí mismo.»[1] Eso dijo G.E. Moore, para acusar después a tal doctrina de «flagrante contradicción».[2] «El principio egoísta», asevera Brian Medlin, «es inconsistente».[3] Con estas acusaciones, Moore y Medlin se erigen en portavoces de una legión de filósofos para quienes el egoísmo no es suficientemente racional. Pero ¿por qué acusar al egoísta? Total, todo lo que hará, en cuanto egoísta, será procurar que le vaya lo mejor posible; y ese propósito, si bien puede no resultarnos del todo atractivo, cae claramente dentro de los confines de la concepción de la racionalidad práctica como maximización de la utilidad, típica de la economía —una concepción a la que difícilmente imputaríamos contradicción o inconsistencia—.[4] Mas los filósofos han sido bendecidos con el talento y el deseo de encontrar paradojas allí don-

1. Moore, G.E., *Principia Ethica*, Cambridge, Cambridge University Press, 1903, pág. 96. Trad. cast.: *Principia Ethica*, México, UNAM, 1983.

2. Ibid., pág. 102.

3. Medlin, B., «Ultimate Principles and Ethical Egoism», *Australasian Journal of Philosophy*, 35 (1957), págs. 111-118.

4. Hay que recordar que el objetivo del egoísta no viene determinado por la concepción maximizadora de la racionalidad práctica. Según esta concepción, es racional actuar con el mayor éxito posible —maximizar cierta medida definida sobre los resultados posibles de las propias acciones—. El requisito de la maximización no determina las características de la medida que ha de ser maximizada. Lo que el egoísta añade a la idea de actuar con el mayor éxito posible es la especificación de que la medida se refiera a uno mismo, de modo que su principio consiste en que le vaya lo mejor posible *a él mismo*.

de otros mortales ni siquiera las sospechan, porque bien pensado ¿qué problema *racional* podría plantear el intento de que a uno le vaya lo mejor posible?

Como filósofo, escondo en la manga algo que, si no verdaderamente paradójico, sí resultará inesperadamente problemático. Pero mis cuestiones sobre el egoísmo no provienen del repertorio filosófico tradicional, sino de la teoría de la decisión racional. En concreto, cuando se coloca al egoísta —que en solitario supera la prueba de la racionalidad— junto a otros egoístas, en una interacción en la que cada cual trata de maximizar su propia felicidad, entonces se presentan fundadas razones para cuestionar la racionalidad del egoísmo. Y esas razones no se refieren al hecho de que el fin del egoísta sea su propia felicidad, sino más bien a su principio maximizador de elección. Algo no funciona en nuestra visión de la racionalidad práctica.

2. Para comprender esto, centrémonos por un momento en la teoría de la decisión racional. Podemos recordar, en primer lugar, una afirmación sentada por John Rawls: «La teoría de la justicia es una parte, quizá la más significativa, de la teoría de la decisión racional».[5] Interpretaré esta afirmación de un modo nada rawlsiano y, para explicar mi interpretación, tendré que aparcar temporalmente el egoísmo; pero antes de hacerlo, quiero destacar una conexión inmediata entre la tesis rawlsiana y nuestro interés sobre la racionalidad del egoísmo. Si la teoría de la justicia es literalmente una parte de la teoría de la decisión racional —hasta el punto de que la justicia es necesaria para tomar una decisión racional—, da la impresión de que, o bien la justicia es compatible con el egoísmo, o el egoísmo no es compatible con la racionalidad. Si lo primero no resultara plausible, entonces la defensa de la racionalidad de la justicia tendría mucho que ver con la argumentación contra el egoísmo. Y ambas dependerían de una adecuada comprensión de la decisión racional.

Antes de concluir estas reflexiones habremos pasado de un argumento contra la racionalidad del egoísmo a un argumento que liga no ya la justicia, sino la moralidad, con la decisión racional. Pero ése será

5. Rawls, J., *A Theory of Justice*, Cambridge (Mass.), Harvard University Press, 1971, pág. 16. Trad. cast.: *Teoría de la justicia*, México, FCE, 1978, pág. 34.

el último eslabón de una larga cadena. Por el momento, analicemos el postulado de Rawls y generalicémoslo en el sentido —que yo creo correcto— de que la teoría moral en su conjunto es parte de la teoría de la elección o decisión racional. He de precisar, sin embargo, que lo que yo creo *no* coincide con lo que cree Rawls.

Yo trato los principios morales como principios *para* la decisión racional. Sostengo que en un tipo de interacción muy generalizado e importante, que denominaré *cooperativa*, un agente racional basaría en un principio moral su elección entre acciones posibles, siempre que esperara que los demás hicieran lo mismo. En la sección quinta de esta parte expondré cuál es ese principio de elección o acción, y en la segunda parte explicaré por qué un agente racional habría de decidir conforme a principios que podemos, con toda propiedad, llamar morales.

Rawls trata los principios de la justicia, no como principios *para* la decisión racional, sino como el objeto *de* una decisión racional,[6] lo cual es muy diferente. Para Rawls, los principios de la justicia determinan la estructura básica de la sociedad. Él pregunta qué principios constitutivos de la sociedad elegiría un individuo racional en la «posición original», tras un velo de ignorancia que le ocultara su identidad y sólo le permitiera saber que es una persona libre e igual. Rawls identifica los principios así elegidos con los principios de la justicia. En eso consiste, según él, la relación de la teoría de la justicia con la teoría de la decisión racional.

Adviértase la diferencia entre nosotros. Rawls pregunta *¿Qué* elegiría un conjunto de agentes racionales tras el velo de ignorancia? Y contesta: elegiría los principios de la justicia. Yo pregunto *¿cómo* decidirían los agentes racionales en la interacción cooperativa? Y contesto: decidirían sobre la base de principios morales. Para Rawls, los principios de la justicia son la solución a un problema concreto de decisión racional.[7] Para mí, los principios morales son empleados por las personas para *solucionar* ciertos problemas de decisión racional. Rawls usa los principios de la decisión racional como herramientas para desarrollar su teoría de la justicia. Yo desarrollo la teoría moral

6. Esta explicación refleja los argumentos de Rawls en *Teoría de la justicia, op. cit.*, secc. 2-4.

7. Véase Rawls, *op.cit.*, secc. 20.

como parte de la teoría de la decisión racional —como parte de la teoría que determina qué principios emplearía un agente racional para tomar sus decisiones.

Existe una segunda diferencia, también importante, entre el intento rawlsiano de usar la decisión racional para definir la justicia y mi intento de desarrollar una teoría de la moralidad como parte de la decisión racional. Y esta diferencia es directamente relevante para mi interés sobre el egoísmo. La Teoría de la Decisión Racional estudia dos formas de acción significativamente diferentes: paramétrica y estratégica.[8] El agente paramétrico decide en un medio que, sea o no conocido en todos sus elementos, supone independiente de su decisión. Su decisión responde a circunstancias que no responderán a su vez (o, al menos, eso se supone), a su condición de decisor. Por el contrario, el agente estratégico elige en un medio sensible a él como decisor. Decide en función de un medio que incluye a otros agentes que tratan a su vez de actuar en función de las decisiones de él. Lo que veremos es que el egoísmo da buenos resultados en la decisión paramétrica, pero no en la estratégica.

A modo de ilustración de la diferencia entre la elección paramétrica y estratégica, consideremos el siguiente ejemplo. Jane tiene que decidir si ir o no a la fiesta de Ana. Ella desea asistir, pero sólo si Brian no lo hace.

Primer caso: Jane espera que Brian vaya a la fiesta siempre que su padre no lo necesite para repartir pizzas, lo cual no tiene nada que ver con Jane. El hecho de que el padre de Brian lo necesite esa noche en el negocio es, para Jane, una circunstancia desconocida, pero independiente de su decisión. Si considera probable que Brian tenga que trabajar, decidirá ir a la fiesta de Ana; y si lo considera improbable, decidirá no ir. Jane está ante un problema de elección paramétrica.

Segundo caso: Brian también tiene que decidir si va o no a la fiesta de Ana. No tiene ningún deseo de ir, excepto si Jane asiste. En este caso, Jane decide basándose en su expectativa sobre la decisión de Brian; y Brian basándose en su expectativa sobre la decisión de Jane.

8. Sobre esta distinción entre acción paramétrica y estratégica, Véase Elster, J., *Ulysses and the Sirens. Studies in Rationality and Irrationality*, Cambridge, Cambridge University Press, 1979; págs. 18-19 y 117-123. Trad. cast.: *Ulises y las sirenas*, México, FCE, 1989.

Así pues, Jane decide basándose en su expectativa de una decisión basada a su vez en una expectativa sobre su decisión. Y Brian lo mismo. Ambos están ante un problema de elección estratégica.

Rawls pone los principios de la justicia en relación a la decisión paramétrica, no estratégica; lo cual parece sorprendente, dado que supone que los principios serían acordados por todas las personas racionales en la posición original. En esa situación, parece lógico pensar que cada uno elegiría ciertos principios en función de lo que decidieran los demás, quienes a su vez estarían decidiendo en función de su decisión. Sin embargo, Rawls subraya que esta apariencia de interacción estratégica es engañosa.[9] Tras el velo de ignorancia, las personas están exactamente en la misma situación, no sólo en sus circunstancias objetivas, sino también subjetivamente, en cuanto que todos ignoran sus capacidades e intereses hasta tal punto que son incapaces de distinguirse a sí mismos de los demás. En tal situación no tienen base alguna para negociar entre sí, y el acuerdo sobre los principios de la justicia puede representarse como la elección de un sólo individuo representativo. El problema de decisión racional que ha de ser resuelto es, en consecuencia, un problema de decisión individual bajo condiciones de extraordinaria incertidumbre. Un problema de elección paramétrica.

Para mí, hay que poner los principios morales en relación con la elección estratégica. Sostendré que los principios morales orientan la elección en la interacción cooperativa, que es aquella en la que cada persona, completamente consciente de sus concretas circunstancias, capacidades e intereses, trata de que sus acciones estén en función de las acciones de otros de un modo que sea beneficioso para todos. El fundamento racional de los principios morales —y, podría añadirse, la irracionalidad de los principios egoístas— surge del análisis de la estructura de tal tipo de interacción.

A la conclusión de nuestro análisis volveré sobre esta diferencia entre Rawls y yo —diferencia que también distingue mi enfoque contractualista del argumento utilitarista de John Harsanyi—.[10] Mi tesis es que hay una profunda incoherencia en el intento de definir los principios morales en función de una decisión paramétrica, ya que la

9. Véase Rawls, *op. cit.*, pág. 139.

10. Véase Harsanyi, John C., *Essays on Ethics, Social Behavior, and Scientific Explanation*, Dordrecht y Boston, Reidel, 1976; caps. 2 y 6, secc. 1-5.

decisión paramétrica no describe en su totalidad la interacción entre seres racionales. La racionalidad estratégica, que refleja mejor esa interacción, no es completamente egoísta, y es lógico que la teoría moral halle su fundamento en el fracaso del egoísmo como principio de racionalidad estratégica.

3. Una vez más, he indicado un punto de destino. Al final de nuestra breve travesía comprenderemos más claramente por qué fracasa el egoísmo y cuál es la relación de la moralidad con la decisión estratégica. Pero, como punto de partida, hemos de explicar la relación del egoísmo con la racionalidad paramétrica.

Supondré que la decisión racional paramétrica puede representarse mediante un modelo maximizador simple. Esto es, que un agente paramétricamente racional se comporta como si estuviera maximizando el valor esperado de una función definida sobre los resultados posibles de sus decisiones. Para que el modelo pueda representar adecuadamente elecciones arriesgadas o inciertas, en las que el agente no conoce el resultado de cada posible decisión sino sólo la probabilidad de cada resultado dentro del conjunto de los que pueden producirse,[11] debe definirse una única función invariable a su transformación lineal positiva, de modo que proporcione una medida de intervalo del valor de los resultados. Un ejemplo familiar de medida de intervalo es la temperatura; el punto cero y la unidad se puede seleccionar arbitrariamente, pero una vez seleccionados, la unidad es constante.

No entraré en el problema de si la maximización capta totalmente la racionalidad de la elección paramétrica. Por el momento, no necesitamos determinar hasta qué punto es cierto que un agente pueda calificarse como racional por el solo hecho de maximizar, al margen de lo que maximice. Así pues, bastará con aceptar que la maximización es uno de los elementos necesarios de la racionalidad paramétrica. Pero he de introducir un importante supuesto adicional: que el valor maximizado por un agente es relativo a él mismo. Si Mary votó por el Partido Republicano y Harry por el Demócrata, podemos suponer que la presidencia republicana tiene mayor valor esperado que la demócrata, como objeto de la elección de Mary, pero menor valor como objeto de la elección de Harry.

11. Hablamos de riesgo si las probabilidades son objetivas, y de incertidumbre si son subjetivas.

Este supuesto de que el valor es relativo y que la decisión está basada en el valor relativo-al-agente sirve tanto para la racionalidad paramétrica como para la estratégica. Aunque concluiremos que la decisión racional estratégica no siempre puede representarse mediante un modelo maximizador simple, sí es posible siempre concebir que son los agentes estratégicamente racionales quienes asignan valores a los resultados posibles. En interacción, las medidas de intervalo definidas por varios agentes sobre los resultados posibles son lógicamente independientes. Puede ocurrir —y de hecho así ocurre— que el resultado más valioso para Brian sea el menos valioso para Jane. Un resultado no tiene un único valor, sino un conjunto de valores, uno por agente —o, mejor, uno por cada una de las personas afectadas— y no hay ninguna relación *a priori* entre los miembros de ese conjunto.

El egoísta, a quien hemos mantenido en un papel secundario durante la discusión sobre la decisión racional, reaparece ahora, sutilmente disfrazado, como una suerte de agente paramétricamente racional. Primero presentamos al egoísta como la persona que persigue la mayor felicidad para sí mismo. Era un maximizador, pero de una cantidad concreta: su propia felicidad. Mas a nuestros efectos, podemos generalizar a partir de esa caracterización y pensar en el egoísta como un maximizador del valor relativo-al-agente que prefiera —quizá su propia felicidad, quizá no—. El egoísta es, pues, la persona cuyos intereses, sean los que sean, no tienen ninguna conexión necesaria con los intereses de los demás, de modo que puede evaluar los estados de cosas sobre la base de sus propios valores, con independencia de los valores de otros. Esta versión genérica nos proporciona una muy débil caracterización del egoísta. De hecho, es demasiado débil, pues admite en las filas egoístas a personas cuyos intereses están orientados a los demás, siempre que esa orientación hacia los demás no esté basada directamente en los intereses de otros. No obstante, basta para nuestro argumento. Por tanto, nuestro egoísta es simplemente un maximizador, o aspirante a maximizador, de valor relativo-al-agente. Con eso satisface la condición necesaria de la racionalidad paramétrica.

Antes de pasar a confrontar al egoísta con los problemas de la interacción, debo reconocer, para ser justo con G.E. Moore, que al introducir la maximización de valor relativo-al-agente ya me he adherido a lo que para él era la contradicción del egoísmo. Para Moore, «el *bien* de [algo] no puede, en ningún sentido posible, ser "privado", o

pertenecerme; del mismo modo que ninguna cosa puede *existir* privadamente o *para* una sola persona».[12] Moore podría admitir que un estado de cosas fomente el bienestar de una persona y no el de otra. Podría admitir que un estado de cosas sea bueno en el sentido de que fomente el bienestar de una persona, y malo en el sentido de que impida el de otra. Pero negaría que un estado de cosas pudiera ser bueno para la persona cuyo bienestar promueva y malo para la persona cuyo bienestar impida. Su tesis es que debe ser absolutamente bueno en tanto fomenta el bienestar de una persona, y absolutamente malo en cuanto impide el de otra.

La posición de Moore podría entenderse como una defensa de la universalidad de las razones para decidir o actuar. Desde ese punto de vista, dados dos estados de cosas cualesquiera p y q, si hay una persona X capaz de decidir entre p y q, y tiene razones para elegir p en vez de q, entonces cualquier persona Y capaz de elegir entre p y q tiene una razón para elegir p en vez de q. Esta posición es adoptada por muchos otros filósofos además de Moore, como R.M. Hare y Thomas Nagel, pero yo, con mi adhesión a la idea del valor relativo-al-agente, me he propuesto ignorarla.[13]

En mi opinión, las razones para elegir tienen, en el mejor de los casos, una universalidad débil. Dados los estados de cosas p y q, si hay una persona X capaz de elegir entre p y q, y tiene una razón para elegir p en vez de q, entonces hay una relación R entre X, p y q, tal que, para cualquier persona Y capaz de elegir entre p y q; a) R no se da necesariamente entre Y, p y q, pero; b) si R se da entre Y, p y q, entonces Y tiene una razón para elegir p en vez de q. Las razones son relativas (válidas para un agente) porque se fundan en una relación entre el agente y los objetos de elección que no existe para todas las personas por el hecho de que se dé para alguna.

Supongamos que Moore y yo estamos de acuerdo en que el incremento de mis perspectivas de supervivencia es una razón para que yo elija que se construya un basurero nuclear en la Antártida en vez de en el condado de Allegheny.[14] Parece que para Moore esto sería un

12. Moore, G.E., *op. cit.*, pág. 99.

13. Véase Hare, R.M., *Moral Thinking*, Oxford, Clarendon, 1981, caps. 5-7; y Nagel, T., *The Possibility of Altruism*, Oxford, Clarendon, 1970, esp. cap. 10.

14. La ciudad de Pittsburgh, donde reside David Gauthier, se encuentra en el condado de Allegheny, Estado de Pennsylvania. (*N. del t.*)

caso particular de la proposición: para todas las personas X e Y, y todos los estados de cosas p y q, si las perspectivas de supervivencia de X son mayores si ocurre p que si ocurre q, entonces Y tiene una razón para elegir p en vez de q. Para mí, el caso ejemplifica esta otra proposición: para toda persona X y estados de cosas p y q, si las perspectivas de supervivencia de X son mayores si p que si q, entonces X tiene una razón para elegir p en vez de q.

Si Moore estuviera en lo cierto y las razones no fueran relativas al agente, entonces la maximización de una medida de resultados posibles relativa-al-agente sería un fundamento irracional para la decisión. Ello supondría el derrumbe no sólo del egoísmo, sino de todo el edificio de la teoría estándar de la decisión racional. Esto permite una fácil refutación del egoísmo, pero a mi juicio es poco convincente. Yo prefiero dar por supuesto que el valor es relativo al agente, y conceder así al egoísmo el primer asalto de la disputa sobre su racionalidad.

4. Podemos ahora pasar a examinar qué sucede cuando el egoísta o, en general, el agente paramétricamente racional, se encuentra en interacción con otros de su misma clase: ¿le conduce su intento de maximizar el valor relativo-al-agente a una contradicción o inconsistencia, o a alguna otra forma de irracionalidad?, ¿le resulta siempre posible poner en práctica su egoísmo? Y, cuando es así, si es que puede serlo en general, ¿resulta siempre racional, o al menos no irracional, hacerlo?

Para resolver, o tratar de resolver, estas cuestiones hemos de fijarnos concretamente en la interacción estratégica, no en la interacción en general. Si el egoísta no tuviera que enfrentarse a problemas estratégicos —problemas en los que trata de responder con su decisión a las decisiones de otros que intentan a su vez decidir en función de la suya— nunca surgirían los problemas que vamos a analizar. En la medida en que la interacción no se conciba en términos estratégicos, el egoísmo parece completamente, incluso paradigmáticamente, racional.

Ésta es una reflexión históricamente importante. Porque la forma de interacción más minuciosamente estudiada, aquella que tiene lugar en el mercado perfectamente competitivo, es paramétrica, no estratégica. Aunque los agentes en el mercado están interactuando con otros de su misma clase, cada uno decide en un medio estable. La empresa trata de maximizar los beneficios, dados los costes (conocidos) de los factores de producción y los precios (también conocidos) de-

terminados por la demanda agregada. El consumidor trata de maximizar el valor del conjunto de bienes y servicios que recibe, dados sus precios. Como las decisiones tienen resultados determinados y conocidos, la interacción de mercado puede representarse mediante un modelo que no requiere una medida de intervalo del valor de esos resultados; es suficiente con una ordenación débil. Un mundo que coincidiese en todos sus detalles con el ideal del mercado perfectamente competitivo no plantearía los problemas que vamos a examinar. El egoísmo es racional en el mercado, como implícitamente reconoció Adam Smith con su doctrina de la «Mano Invisible»; y el atractivo actual del egoísmo no es ajeno al dominio del modelo del mercado en nuestro pensamiento práctico.

Pero no todo comportamiento económico es perfectamente competitivo, y no todo comportamiento es económico. El mercado es un modelo adecuado para un ámbito de la interacción; pero sólo para ese ámbito. Cuando Von Neumann y Morgenstern titularon su libro *Theory of Games and Economic Behavior* (*Teoría de juegos y comportamiento económico*), estaban llamando la atención sobre la —hasta entonces insuficientemente comprendida— dimensión estratégica de la mayoría de interacciones.[15] Y es esta dimensión la que nos interesa, puesto que nuestro análisis se centra en los problemas que surgen al intentar aplicar el modelo maximizador simple de la racionalidad paramétrica a los contextos de decisión estratégica.

Descubriremos que se plantean dos problemas principales. El primero se puede expresar diciendo que el egoísmo es *inconsistente* —o incapaz de proporcionar una orientación consistente para la elección en general—. El segundo se resume en la tesis de que el egoísmo es *contraproducente* (*self-defeating*) —es decir, que los egoístas realizan sus objetivos en mucha menor medida que algunos no egoístas—. La primera imputación, no tiene fácil salida. La segunda es, sin embargo, discutible; y su discusión nos abocará al problema constructivo para el cual nuestro argumento sirve de propedéutica —el desarrollo de la teoría moral como parte de la decisión racional—. Entonces comprenderemos por qué el agente estratégicamente racional debe ser (o al menos debe convertirse en) un agente moral.

15. La obra de Von Neumann y Morgenstern es el trabajo seminal a partir del que se han desarrollado los estudios sobre racionalidad estratégica.

Vamos a ejemplificar ahora las dos imputaciones que hemos de considerar; luego, dedicaré el resto de esta parte a examinar la inconsistencia, dejando para la segunda la «contraproductividad» (*self-defeatingness*).

La objeción de que el egoísmo es inconsistente puede ilustrarse con ayuda de nuestro ejemplo inicial de decisión estratégica. Jane y Brian deben decidir si van a la fiesta de Ana. Suponemos que cada uno tiene dos, y sólo dos, posibilidades: ir o no ir. Si ambos deciden ir, entonces Jane tomó la decisión equivocada; ella quiere ir a la fiesta, pero sólo si Brian no asiste. Si ninguno de los dos decide ir a la fiesta, entonces Jane tomó de nuevo una decisión equivocada; pues, dado que Brian no asiste, ella habría preferido ir. Si uno decide ir y el otro no ir, entonces es Brian quien tomó la decisión equivocada; pues él quiere ir a la fiesta, pero sólo si Jane asiste. Decidan lo que decidan Jane y Brian, uno de los dos no logra maximizar su valor. Por lo tanto, uno no logrará satisfacer las demandas del egoísmo. Como este fracaso es inevitable, puede decirse que es imposible que se satisfagan en general las demandas del egoísmo. En conclusión: el egoísmo —y toda maximización de valor relativo-al-agente— es inconsistente.

La imputación de que el egoísmo es contraproducente (*self-defeating*) se puede aclarar con un ejemplo familiar para los teóricos de juegos desde hace tiempo, y ya ampliamente divulgado entre los filósofos: el Dilema del Prisionero. Jack y Zack son presos acusados de un delito grave; cada uno tiene que decidir entre una confesión que implicaría al otro, o no confesar. Si sólo uno confiesa, las leyes del Estado le conceden, como premio por esclarecer el crimen, una sentencia benigna, mientras que el acusado recibe la máxima pena. Si ambos confiesan, los dos reciben una dura sentencia, pero sin llegar a la máxima pena. Si ninguno de los dos confiesa, ambos serán convictos de un delito menor, cuya sentencia es algo mayor que la que premia la confesión. Jack calcula que, si Zack confiesa, él puede evitar la máxima pena confesando; y si Zack no confiesa, él puede conseguir la pena más liviana confesando. Haga lo haga Zack, a Jack le va mejor si confiesa. Por supuesto, Zack razona de modo análogo. Dado que ninguno puede influir con su decisión en la decisión del otro, es mejor para cada uno confesar, haga lo que haga el cómplice. Jack y Zack maximizan su valor confesando. Como consecuencia, cada uno recibe una dura sentencia. Si ninguno hubiera confesado, ambos habrían

recibido una sentencia más benigna. Jack y Zack han cumplido las demandas del egoísmo y han alcanzado un resultado costoso para ambos. Por tanto, no hay razón para imponer esas demandas en todo caso. El egoísmo es contraproducente (*self-defeating*).

5. Quizá parezca excesivo acusar al egoísmo de inconsistencia y añadir luego que es contraproducente. Si el principio egoísta impone demandas que no pueden ser satisfechas simultáneamente, no parece tener sentido preguntarse si cumplir con esas demandas es, además, contraproducente. Lo cierto es que la inconsistencia no afecta a todas las situaciones en que las personas tratan de actuar de modo egoísta. El fracaso del egoísmo como guía para la acción se produce sólo en algunas interacciones, como la de Jane y Brian.

La naturaleza de este fracaso debe quedar clara. Que uno de los dos (Jane o Brian) no alcance su resultado preferido, no significa que su decisión haya sido incorrecta. Si Brian decide ir a la fiesta de Ana, Jane no puede alcanzar su resultado preferido (que es disfrutar de la fiesta sin Brian). Si Jane decide quedarse en casa, es Brian quien no puede lograr el resultado que prefiere. En estos casos, el resultado favorito de una persona es eliminado por la elección de otra. Por tanto, la imposibilidad de alcanzarlo no muestra necesariamente que la decisión sea equivocada y, en consecuencia, no plantea problema alguno para el egoísmo. Se podría pensar que el egoísmo es inconsistente porque los egoístas tienen objetivos incompatibles, de modo que todos no pueden lograrlos a la vez. Sin embargo, la mera existencia de objetivos incompatibles no impide que cada individuo obtenga el mejor resultado posible para él —entendiendo que «lo posible» vendrá determinado en parte por las decisiones de los demás—. La situación que estamos analizando presenta, sin embargo, un caso en que uno de los dos no logra alcanzar su mejor resultado posible dada la decisión del otro; y es este fracaso —el no lograr elegir lo que maximizaría el valor propio dadas las posibilidades que la decisión del otro permite— el que hace surgir el reproche de que el egoísmo es incapaz de ofrecer una guía consistente con carácter general.

Así pues, la inconsistencia del egoísmo se plantea del siguiente modo: el egoísta debería, en la interacción estratégica, maximizar el valor relativo-al-agente. ¿Qué forma de actuar debería adoptar el agente para cumplir ese objetivo? Parece que tendría que elegir siempre aquello que maximizara el propio valor, dadas las decisiones de

los demás. Ahora bien, nuestro ejemplo muestra que a veces es imposible que todos elijan de ese modo. *Cualquier* persona puede hacerlo: dadas las decisiones de otros, cualquiera tiene una alternativa maximizadora. Pero no *todas* pueden hacerlo. No todos pueden ser siempre maximizadores de valor relativo-al-agente en la interacción estratégica. Y como el egoísmo impone a todos ese objetivo, es inconsistente.

Pero hemos ido demasiado aprisa. En un mundo arriesgado e incierto, ni siquiera el agente paramétricamente racional puede estar seguro de maximizar el valor relativo-al-agente. Puede, como mucho, maximizar el valor relativo-al-agente *esperado*, dada su estimación de la probabilidad de que las circunstancias sean unas u otras. Análogamente, el agente estratégicamente racional debe contentarse con maximizar el valor esperado. Esto se traduce en que no elegirá necesariamente lo que maximiza su valor dadas las decisiones de los demás, sino lo que maximiza su valor dadas las decisiones esperadas de los demás. Si Jane juzga improbable que Brian decida ir a la fiesta, tal vez ella maximice su valor esperado decidiendo ir. Si Brian cree probable que Jane decida ir a la fiesta, quizás él maximice su valor esperado optando por ir. El resultado será decepcionante para Jane, pero su decisión satisfizo formalmente las demandas del egoísmo. No hay razón, por tanto, para negar que todos puedan ser maximizadores del valor relativo-al-agente esperado, aunque el resultado final decepcione a algunos.

Mas esta réplica procede también demasiado rápido. Supongamos que Jane y Brian saben que el otro aspira a maximizar su propio valor. Ello implica que, para maximizar el valor esperado, cada uno debe elegir sobre la base de la expectativa de la decisión que el otro tomará basándose en su decisión. Jane decidirá ir a la fiesta si espera que Brian decida no ir, con base en la expectativa de que ella no irá. Jane decidirá quedarse en casa si espera que Brian decida ir a la fiesta, por estar convencido de que ella ha decidido también asistir. Cualquier decisión de Jane que maximice su valor esperado, tendrá su causa en la suposición de que Brian se ha formado una expectativa incorrecta sobre su decisión. Análogamente, Brian debería basar su decisión en la suposición de que Jane ha formado una expectativa equivocada sobre la decisión tomada por él.

Esto nos permite dar una explicación más exacta de lo que pare-

ce ser la causa de la inconsistencia del egoísmo. En la decisión estratégica, el egoísta debe maximizar el valor relativo-al-agente esperado. En consecuencia, debe intentar maximizar su valor, dadas las decisiones esperadas de otros que también tratan de maximizar su valor dadas las decisiones de los demás, incluido el agente que está deliberando. Pero en relación con esta consecuencia, hay tres proposiciones que no pueden ser verdaderas a la vez:

> 1. Un egoísta siempre decide maximizar su valor dadas las decisiones esperadas de los demás.
> 2. Un egoísta siempre espera que otros egoístas decidan maximizar su valor propio dadas las decisiones esperadas de los demás (el deliberador incluido).
> 3. Satisfacer 1 y 2 nunca exige al egoísta suponer que las expectativas de otros egoístas son incorrectas.

El problema del egoísmo radica, por tanto, en que —en situaciones como la de Jane y Brian—, para suponer que todos deciden maximizar el valor relativo-al-agente dada su expectativa sobre las decisiones de los demás, el egoísta tiene necesariamente que atribuir a otros expectativas incorrectas.

Vamos a introducir cierta terminología para expresar con precisión lo que acabamos de exponer. Una acción que maximiza el valor del agente en interacción con otros es una *mejor respuesta* a las acciones de los otros. Un egoísta elige la mejor respuesta esperada. A veces, no existe ningún conjunto de acciones —una para cada agente— tal que todos sus elementos sean una mejor respuesta a los elementos restantes. En la terminología de la Teoría de Juegos, un conjunto de mejores respuestas mutuas es un *equilibrio de Nash*;[16] hay situaciones en las que ningún conjunto de acciones es un equilibrio de Nash. En tales situaciones, los egoístas sólo pueden elegir mejores respuestas esperadas si algunos se forman expectativas erróneas. La existencia de un conjunto de acciones en equilibrio es, por tanto, una condición necesaria para que una decisión egoísta bien informada tenga éxito.

16. El término «Equilibrio de Nash» hace referencia a John F. Nash, autor de las ideas fundamentales sobre el equilibrio en la interacción estratégica, que discutiremos en la sección siguiente.

Digamos que un *principio de decisión* es una función que selecciona subconjuntos de acciones pertenecientes a conjuntos de acciones alternativas (dado un conjunto cualquiera C, el subconjunto correspondiente es denominado *conjunto de decisión* «D(C)»). Un principio es *completo* en un dominio si y sólo si asigna algún subconjunto no vacío para cada miembro (conjunto) del dominio. Un principio es *egoísta* sólo si para cada conjunto C selecciona un subconjunto D(C) cuyos miembros maximizan alguna medida definida sobre C. En una comunidad de egoístas suficientemente informados, el principio que determine una decisión para cada persona implicada en una interacción, debe seleccionar decisiones que maximicen el valor para cada persona, dadas las demás decisiones determinadas por dicho principio. En otras palabras, un principio que incluya en su dominio todos los conjuntos de acciones alternativas que forman la interacción, debe seleccionar, para cada conjunto, un subconjunto cuyos miembros sean acciones en equilibrio de Nash. Puesto que para algunas interacciones no hay ningún conjunto en equilibrio, no puede haber ningún principio de decisión egoísta definido sobre el dominio de todos los conjuntos de acciones alternativas en todas las interacciones posibles: no existe un principio de decisión egoísta que pueda ser completo para toda interacción estratégica. Éste es el sentido preciso de la acusación de que el egoísmo es inconsistente.

6. El recurso a la teoría de la decisión racional da lugar a complejidades inesperadas en nuestra pesquisa sobre la consistencia del egoísmo. Y eso que sólo hemos concluido el primer asalto de nuestra disputa; el segundo comienza ahora, al pasar de las *acciones* a las *estrategias*. Una estrategia es una lotería, o distribución de probabilidades, entre acciones posibles. Hasta este momento, hemos supuesto que cada agente decide entre acciones; ahora vamos a ampliar el espacio de decisiones, y pensaremos que cada agente puede elegir entre estrategias. De hecho, elegir una acción equivale a elegir una estrategia que asigna a esa acción una probabilidad 1 y a las restantes acciones alternativas una probabilidad 0. Tal tipo de estrategia se denomina *pura*. Pero hay innumerables estrategias *mixtas*, que asignan una probabilidad positiva a cada una de las dos o más acciones posibles.

Hemos supuesto que las decisiones de cada agente pueden interpretarse como el intento de maximizar el valor, o el valor esperado, de una función que ordena, sobre una escala de intervalo, los resulta-

dos posibles. El valor asignado a cada acción es la suma ponderada de sus resultados posibles, donde el peso de cada resultado depende de la probabilidad de que se produzca de hecho una vez realizada la acción. El valor asignado a cada *estrategia* será, según la misma lógica, la suma ponderada de los valores de las acciones que la componen, donde el peso de cada acción viene determinado por la probabilidad que esa estrategia concreta le asigna. Y, como en el caso de las acciones, supondremos que, al decidir entre las estrategias posibles, el egoísta tratará de maximizar un valor relativo-al-agente.

Por sorprendente que parezca, esta suposición nos permite liberar al egoísta del reproche de dejarse guiar por un principio de decisión inconsistente. Ello es debido a que, hace más de treinta años, John F. Nash probó que, en cualquier interacción entre un número finito de personas, cada una de ellas con un número finito de acciones o estrategias puras, hay al menos un conjunto de estrategias en equilibrio.[17] En otras palabras, existe al menos un conjunto de estrategias, una por agente, cada una de las cuales es una mejor respuesta a las demás del mismo conjunto. Y la existencia de tal conjunto en equilibrio satisface la demanda de que la decisión egoísta sea maximizadora y esté basada en expectativas correctas.

Para aplicar el teorema de Nash sobre la existencia de un conjunto de estrategias en equilibrio a la resolución del problema de decisión de Jane y Brian, tendríamos que establecer, para cada uno, una medida de intervalo de los resultados posibles, y resolver luego el problema matemático resultante. En vez de hacer eso, vamos a desarrollar intuitivamente la idea de la determinación de un par de estrategias cada una de las cuales es la mejor respuesta a la otra. Supongamos que Jane no confía en poder ocultar su decisión estratégica a Brian. Se teme que, si adopta una estrategia en la que ir a la fiesta de Ana tiene una alta probabilidad, Brian responderá yendo él mismo a la fiesta, con lo cual el resultado probable no será el deseado por ella. Y si adoptara una estrategia en la que ir a la fiesta tiene muy baja probabilidad, Brian respondería quedándose en casa, de modo que de nuevo el resultado más probable será el menos apetecible para ella. ¿Qué puede hacer, entonces? Necesita una estrategia ante la que

17. Nash, J.F., «Noncooperative Games», *Annals of Mathematics*, 54 (1951), págs. 286-295.

Brian sea indiferente entre ir o no ir a la fiesta; que le proporcione el mismo valor esperado en un caso o en otro. Análogamente, Brian necesita una estrategia que deje a Jane indiferente entre las dos opciones. Si Brian es indiferente entre sus estrategias posibles, entonces adoptará cualquier estrategia como mejor respuesta; igualmente, si Jane es indiferente entre sus estrategias posibles, entonces cualquiera de esas estrategias es su mejor respuesta. Por lo tanto, si ambos eligen la estrategia que deje al otro indiferente, cada una será la mejor respuesta a la otra, de modo que el par constituirá un conjunto en equilibrio de Nash.

No es que Jane se forme una expectativa sobre la decisión de Brian y decida cuál es su mejor respuesta a ella. Más bien elegirá una estrategia que deje a Brian sin elección posible entre sus respuestas disponibles. Y Brian hará lo mismo: elegirá una estrategia que deje a Jane sin elección posible. En situaciones del tipo de la que estamos considerando, siempre existe una estrategia que deje al otro indiferente, y esas estrategias ¡son, cada una, la mejor respuesta a la otra; con lo cual, parece que sí es posible que todos los agentes lleven a cabo una decisión egoísta.

Suponiendo valores razonables (para Jane y Brian) de los resultados de las decisiones, se llega a la conclusión de que Jane debería elegir una estrategia mixta consistente en ir a la fiesta con una probabilidad de 2/7, y no ir con una probabilidad de 5/7, mientras que Brian debería elegir una estrategia mixta compuesta por 3/5 de probabilidad de ir a la fiesta y 2/5 de no ir.[18] Estas estrategias constituyen el único equilibrio de Nash. Calcular una de estas estrategias presupone un conocimiento que permite calcular también la otra.

18. Estas estrategias mixtas estarían en equilibrio en el siguiente caso. Asignando arbitrariamente el valor 1 al resultado favorito del agente y 0 al resultado menos deseado, encontramos que una medida de intervalo de las preferencias de Jane daría el valor 1 al resultado de ir a la fiesta de Ana si Brian no va, 1/2 al resultado de no ir si Brian va, 1/4 al de no ir si Brian tampoco va, y 0 al de ir y encontrar allí a Brian. Una medida de intervalo de las preferencias de Brian asigna 1 al resultado de ir a la fiesta si Jane va también, 1/2 al de no ir si Jane no va, 1/10 al de ir si Jane no va, y 0 al de no ir si Jane asiste. La estrategia mixta de Jane dejaría a Brian con una utilidad esperada de 5/14, hiciera lo que hiciera; y la estrategia mixta de Brian proporcionaría a Jane, en cualquier caso, una utilidad esperada de 2/5. Hay que advertir que los valores de utilidad para Jane y Brian *no* son comparables interpersonalmente; es ilegítimo inferir que Jane sale mejor parada de esta situación por el hecho de que 2/5 sea mayor que 5/14.

En ese caso, ninguno de los dos necesitaría molestarse en ocultar su estrategia. Claro que, llegado el momento, cada uno tendría que determinar qué acción física realizar —sin duda empleando una eficaz máquina portátil de selección aleatoria de resultados, programable para cualquier lotería—. Hay que suponer que el resultado de esa determinación es desconocido para el otro hasta que la acción es llevada a cabo. Al imaginar que cada uno elige una estrategia, estamos ya suponiendo que ambos piensan en la decisión del otro sólo en términos de estrategia, sin formarse expectativas más determinadas sobre su acción. Si Brian pudiera saber que en la máquina de decisión aleatoria de Jane salió «ir a la fiesta», entonces él ya no consultaría su propia máquina, sino que se dirigiría inmediatamente a casa de Ana.

No tenemos ninguna razón para suponer que Jane y Brian tienen de hecho la información sobre los valores del otro que sería necesaria para calcular las estrategias en equilibrio. Y eso que se trata de una interacción muy simple. En situaciones más complejas, el procedimiento para determinar las estrategias en equilibrio puede ser más complicado, incluso aunque se disponga de la información necesaria para calcularlas. La prueba de Nash nos asegura que debe haber al menos un conjunto de estrategias en equilibrio, pero muy bien puede suceder que este conocimiento no tenga ninguna aplicación práctica. Por tanto, nada definitivo puede afirmarse sobre la capacidad de los egoístas reales para elegir siempre estrategias que sean mejores respuestas mutuas. Pero una cosa es reconocer que los egoístas pueden no lograr su objetivo, y otra muy distinta demostrar que *necesariamente* fracasan en su intento. Existe un principio de decisión entre estrategias cuyo dominio incluye todos los conjuntos de estrategias alternativas que componen la interacción y cuyos valores son subconjuntos en los que todos los elementos son estrategias en equilibrio mutuo. Y siendo esto así, no se puede decir que la tarea de los egoístas sea irrealizable por principio.

Para evitar posibles malentendidos, hay que advertir que la consistencia estratégica del egoísmo no afecta en modo alguno a la imposibilidad ocasional de seleccionar acciones que cumplan el requisito egoísta. Cuando Jane y Brian actúan de hecho, y descubren la acción realizada por cada uno, entonces uno de ellos no maximizará el valor dado al comportamiento del otro. Ascender al nivel estratégico no am-

plía las posibilidades reales de acción y, por lo tanto, no afecta a la imposibilidad de que Jane y Brian logren su objetivo maximizador en términos de sus acciones cuando su información es correcta. Pero si cada uno selecciona una estrategia que es la mejor respuesta a la elegida por el otro, ambos sabrán que —cualquiera que sea el resultado final— han maximizado su valor esperado. Ninguno pensará que su decisión ha sido fallida *como decisión*.

Antes de concluir esta fase de nuestro argumento, hemos de admitir que no hemos mostrado la existencia de un principio de decisión egoísta válido para todos los conjuntos de estrategias alternativas de todas las interacciones posibles en su dominio. Todo lo que hemos mostrado es la posibilidad de satisfacer el requisito de que las estrategias seleccionadas por el principio en cualquier interacción constituyan un equilibrio de Nash. Pero no se debe suponer que el único requisito de un principio para la decisión egoísta es que seleccione una estrategia —cualquiera— perteneciente al conjunto de estrategias en equilibrio. Esto bastaría, ciertamente, en la situación simple que hemos analizado, en la que sólo hay un par de estrategias en equilibrio; pero en otras situaciones más complejas, puede haber muchos conjuntos formados por estrategias en equilibrio, de tal modo que una estrategia que es la mejor respuesta a las demás de su conjunto, no lo es ante las pertenecientes a un conjunto distinto. Piénsese, por ejemplo, en una situación en la que varias personas desean reunirse, pero les da igual el lugar (entre los varios posibles). Si todos deciden ir al mismo lugar, entonces la acción de cada uno es la mejor respuesta a las de los demás; pero si deciden ir a lugares diferentes, las acciones no son mejores respuestas. No hemos analizado cómo se las ingeniarían los egoístas, abrumados con tantas posibilidades, para decidir entre diferentes conjuntos de acciones o estrategias en equilibrio. Así, esta fase de nuestro argumento concluye simplemente con el dictamen de que la acusación de inconsistencia contra el egoísmo no ha sido probada.

7. El análisis exhaustivo de los problemas que se plantean a los egoístas por la existencia en algunas situaciones de varios conjuntos de acciones o estrategias en equilibrio, queda fuera del alcance de la presente investigación. Me centraré sólo en uno de esos problemas, planteado por el razonable requisito de que los egoístas coordinen sus decisiones para producir un equilibrio superior, al menos en las

situaciones en que tal equilibrio superior exista y la falta de coordinación perjudique a todos.[19]

Comencemos considerando un juego simple. Consiste en dos jugadores a los que se da una moneda, y tienen que elegir entre mostrar la cara o la cruz. No se permite que se comuniquen; cada uno debe decidir sin saber qué hará el otro. Si ambos muestran lo mismo, ambos ganan una suma de dinero, pero la suma es mayor si ambos muestran cara. Si uno muestra cara y el otro cruz, ambos pierden una cantidad de dinero. En este juego hay dos pares de acciones en equilibrio de Nash: que ambos muestren cara y que ambos muestren cruz. Pero el primer par es un equilibrio superior, que domina al segundo, puesto que el resultado de mostrar cara tiene mayor valor para ambos que el de mostrar cruz. Es evidente que un principio de elección egoísta tendría que tener esto en cuenta. Nuestra sugerencia inicial sería que el principio en cuestión debe exigir que se elijan las estrategias que aseguren la coordinación en el equilibrio superior, siempre que haya un equilibrio superior y sea accesible (ignoraremos aquí ciertas condiciones necesarias de accesibilidad).[20] Así, es de suponer que en este juego los egoístas racionales elegirían cara.

Una variante de nuestro juego sugiere, sin embargo, que la exigencia de coordinación propuesta quizá sea demasiado fuerte para los egoístas. Supongamos que cada jugador gana una cantidad de dinero si el otro muestra cara y pierde la misma cantidad si el otro muestra cruz. En este juego, todos los pares de estrategias están en equilibrio, puesto que ambos jugadores son totalmente indiferentes entre sus dos opciones; lo que cada uno gana depende de lo que el

19. El problema discutido en esta sección es esencialmente el mismo que analizo en «The Impossibility of Rational Egoism», *Journal of Philosophy* 4 (1975), págs. 411-433. Aquel artículo examinaba algunos detalles que no trato aquí; por contra, no trataba de modo directo el tema de la aplicación consistente de un principio de decisión a una interacción y sus sub-interacciones.

20. Piénsese en una situación con tres resultados posibles, cada uno de ellos derivado de un conjunto de estrategias en equilibrio. Llamemosles p, q y r, y supongamos que p y q son indiferentes (desde el punto de vista de cada individuo), pero superiores a r. Si la comunicación es imposible entre los jugadores, y ni p ni q poseen ninguna característica distintiva natural, puede que la única coordinación posible sea en el equilibrio inferior r, porque el hecho de ser inferior lo distingue, mientras que nada distingue p y q. En este caso, p y q son inaccesibles de hecho; en ausencia de comunicación, no hay modo de determinar ninguno de ellos como el objetivo de la coordinación.

otro elija. Hay un solo equilibrio que es superior a todos los demás (el que consiste en que ambos jueguen cara). Pero la exigencia de coordinarse en ese equilibrio no tiene una motivación egoísta. No hay ningún incentivo para mostrar cara, puesto que mostrar cruz no reduce el valor esperado ni implica renunciar al equilibrio. En este juego no tenemos razón alguna para suponer que los egoístas racionales elegirían cara en vez de cruz.

La explicación intuitiva es que la coordinación en estrategias pertenecientes a equilibrios superiores debe considerarse una exigencia egoísta sólo si la defección de tal coordinación reduce el valor esperado del desertor. Me ahorraré, sin embargo, la formulación precisa de esta condición, ya que en la situación que vamos a considerar seguidamente (que plantea un problema para la consistencia del egoísmo) quedará clara la motivación egoísta de la coordinación.

Consideremos un juego más complejo, en el que tres jugadores, A, B y C, han de mostrar, sin comunicarse entre ellos, la cara o la cruz de una moneda. Los valores, o pagos, de los resultados de las diferentes combinaciones de acciones son los siguientes:

| *Acción* | | | *Pago* (= estrategia) | | |
|---|---|---|---|---|---|
| A | B | C | A | B | C |
| Cara | Cara | Cara | $1.00 | $1.00 | $1.00 |
| Cara | Cara | Cruz | $1.50 | $1.50 | 0 |
| Cara | Cruz | Cara | –$1.50 | –$1.50 | 0 |
| Cruz | Cara | Cara | –$1.50 | –$1.50 | 0 |
| Cara | Cruz | Cruz | $1.50 | –$3.00 | $1.00 |
| Cruz | Cara | Cruz | –$3.00 | –$1.50 | $1.00 |
| Cruz | Cruz | Cara | $1.50 | $1.50 | 0 |
| Cruz | Cruz | Cruz | $1.00 | $1.00 | $1.00 |

En este juego hay un único conjunto de estrategias en equilibrio: todos los jugadores muestran cara. Esto se puede verificar fácilmente; jugar cara es la mejor respuesta de cada jugador si los demás muestran cara, de modo que ese conjunto está en equilibrio. Es único ya que, dado cualquier otro conjunto, al menos un jugador puede mejo-

rar su pago cambiando su respuesta, de modo que los demás conjuntos de estrategias no son equilibrios de Nash.

Supongamos que A espera que C juegue cara. Esto es lo lógico, ya que A piensa que, si todas las expectativas son correctas y cada decisión es la mejor respuesta a las demás, las estrategias estarán en equilibrio, y jugar cara es el único equilibrio. Pero se da cuenta de que *si* C juega cara, entonces B y él obtendrían un resultado mejor jugando cruz (ganarían $1.50 cada uno, en vez de sólo un dólar). Si tomamos la decisión de C como dada (debido al requisito del equilibrio) y nos centramos únicamente en la interacción entre A y B, los pagos asociados a los resultados posibles son los que se muestran en la siguiente tabla:

| *Acción* | | *Pago* | |
|---|---|---|---|
| A | B | A | B |
| Cara | Cara | $1.00 | $1.00 |
| Cara | Cruz | –$1.50 | –$1.50 |
| Cruz | Cara | –$1.50 | –$1.50 |
| Cruz | Cruz | $1.50 | $1.50 |

En este subjuego hay dos pares de estrategias en equilibrio —ambos jugadores juegan cara o ambos cruz—, de las que la última domina a la primera (es un equilibrio superior). La coordinación en este equilibrio tiene una motivación egoísta: ambos salen ganando con la coordinación y ambos pierden si se apartan de ella. Si A espera que el razonamiento de B sea análogo al suyo, concluirá que, dado que es plausible esperar que C juegue cara, él debería mostrar cruz, con la expectativa fundada de que B hará lo mismo.

Sin embargo, la deliberación de C no tiene por qué terminar al darse cuenta de que la condición del equilibrio le determina a elegir cara. Si anticipa el razonamiento de A y B, que les lleva a coordinarse en mostrar cruz, advertirá que él mismo debería jugar cruz. Ésa es la mejor respuesta si espera que los otros elijan cruz (le permite mantener su ganancia de un dólar, en vez de perderlo en beneficio de A y B). Ahora bien, si A y B anticipan esta deliberación adicional de C, lo que deben hacer es coordinarse en jugar cara; pues si esperan que C juegue cruz, el par de estrategias en que ellos eligen cara domina al

par en que ambos eligen cruz (les permite recapturar el beneficio de C). Y si C anticipa esta deliberación ulterior de A y B, supondrá que ambos van a elegir cara, y él hará lo propio. Con lo que regresamos al punto de partida del círculo que hemos trazado.

El argumento implícito en el análisis de este juego sugiere una tesis sobre el sentido en que un principio de decisión ha de ser consistente para poder ser empleado con éxito. Supongamos que un principio incluye en su dominio todos los conjuntos de estrategias que forman una interacción, de modo que determina para cada agente un subconjunto de sus estrategias: su conjunto de decisión. Supongamos que cada subconjunto contiene sólo una estrategia (lo cual sucederá si el principio satisface la condición de equilibrio y la situación tiene sólo un equilibrio de Nash). El principio requiere que cada agente elija la única estrategia en su conjunto de decisión. Tomando la decisión de cualquier agente, así determinada, como una circunstancia fija, podemos aplicar el principio a la interacción reducida entre los restantes agentes. Pues bien, nuestra tesis es que si el principio es consistente, debe asignar a cada agente de los que componen la interacción reducida, un conjunto de decisión que contenga la estrategia presente en el conjunto de decisión inicial. El principio debe ofrecer una guía consistente, tanto si un agente lo aplica directamente a su elección de estrategia en la interacción global, como si, dando por supuesto que algunos otros se conformarán a él, lo aplica a su elección de estrategia en la subinteracción resultante. Un principio que dijera: «Todos deben jugar cara, pero si el agente C juega cara, entonces todos los demás deben jugar cruz», sería inconsistente.

Si aceptamos esta interpretación de la consistencia, entonces ningún principio egoísta puede ser a la vez completo y consistente. Un principio egoísta debe satisfacer la condición del equilibrio —que exige que las estrategias elegidas en una interacción sean mejores respuestas cada una a las demás— y la condición de la coordinación —con su demanda de que las estrategias elegidas produzcan el equilibrio superior, en los casos en que exista y la defección sea costosa para el desertor—. Para ser completo, un principio de decisión debe asignar un conjunto de decisión no vacío para cada conjunto de estrategias en cada interacción posible. Por último, un principio de decisión consistente debe poder determinar conjuntos de decisión tanto para una interacción en su totalidad, como para cualquier subinteracción

que resulte de tomar la decisión de alguno de sus participantes como dada; y debe hacerlo de tal modo que esos conjuntos sean compatibles entre sí. Pues bien, nuestro juego de las monedas muestra que el egoísmo, la completitud y la consistencia no se pueden dar a la vez.

Los egoístas tratan de responder a las acciones de los demás de un modo que les sea individualmente ventajoso, y tratan también de coordinar sus acciones cuando es beneficioso. Está demostrado que estos objetivos son incompatibles en situaciones como la del juego analizado. Desde luego que no pasa nada por no lograr un objetivo. Pero lo que hemos mostrado es que en ciertas situaciones, algunos agentes no pueden alcanzar su mejor resultado posible, dadas las decisiones de los demás. Si A, B y C no juegan cara todos ellos, entonces al menos uno de ellos podría obtener un resultado mejor, dadas las posibilidades que las decisiones de los demás dejan abiertas. Y si A, B y C juegan todos ellos cara, entonces A y B podrían coordinarse en un resultado mejor para ellos. En tal situación, el egoísmo impone condiciones inconsistentes y, por tanto, fracasa.

8. Ante esto, ¿qué puede hacer el egoísta? Desde luego, siempre podrá tomar su mejor decisión. Sin embargo, ya sabemos que esta respuesta es engañosa. Hemos descubierto situaciones en las que no todos *pueden* tomar su mejor decisión. *Cualquiera* puede hacerlo, pero *no todos*. Si preguntamos qué pueden hacer *los egoístas*, no cabe responder que *ellos* pueden tomar su mejor decisión.

Cambiemos, pues, la cuestión. ¿Qué pueden elegir los egoístas? Si suponemos que el objeto de elección son las estrategias, nuestra respuesta sería: pueden elegir la mejor posible. Sin embargo, sabemos que esta respuesta también es engañosa. Hemos descubierto situaciones en las que todos pueden elegir la mejor estrategia, dadas las elecciones de los demás; pero sólo si algunos dejan pasar la oportunidad de coordinar sus elecciones en lo que es, para ellos, la mejor estrategia. Y si éstos lograran coordinarse, entonces algún individuo no podría elegir su mejor estrategia. Individualmente, los egoístas sí pueden elegir su mejor estrategia; pero coordinadamente (valga la expresión) no todos pueden.

Enfrentado a las complejidades de la interacción estratégica, el egoísta perderá pronto la ingenua esperanza de formular un principio de decisión consistente y completo, acorde con las condiciones implícitas en su actitud egoísta. El problema es más arduo de lo que han

reconocido tanto los críticos como los defensores del egoísmo. Quizá, pues, la primera lección para el aspirante a egoísta sería confiar menos en las palabras de los filósofos y prestar mayor atención a las estructuras de interacción descubiertas por los teóricos de juegos. Lo malo es que el mensaje que trae nuestra investigación basada en la teoría de juegos no es precisamente alentador: cualquiera puede lograr su objetivo egoísta, pero no todos.

Claro que quizá no habría que exagerar la importancia de la consternación y perplejidad que esta conclusión causará sin duda en el aspirante a egoísta. Lo que es demostradamente imposible es, sencillamente, imposible. El egoísta puede tomar su mejor decisión. Es cierto que las estructuras de interacción restringen las decisiones posibles de modo inesperado, pero eso no contrae ni expande el horizonte real de decisiones y acciones que los egoístas afrontaban desde un principio.

Tal actitud minimizaría demasiado fácilmente la relevancia de nuestro argumento. Los egoístas —y no sólo ellos, sino todos los maximizadores de valor relativo-al-agente— han ignorado, por regla general, la estructura del problema en que se hallan. Han reconocido las dificultades derivadas de la incompatibilidad entre los valores que profesan; la cual se ha interpretado a veces como irracionalidad, y otras simplemente como el origen de una inevitable frustración. Pero no han reconocido las limitaciones para tomar su mejor decisión, o elegir su mejor estrategia. No han reconocido que el hecho mismo de acertar en las expectativas sobre las decisiones de los demás participantes en la interacción, asegura a veces que algunos no puedan *hacer* lo que maximizaría su valor, dadas las posibilidades que las decisiones esperadas de los demás dejan, de hecho, abiertas. No han reconocido que una completa consciencia de las posibilidades de la coordinación tiene como consecuencia necesaria que, o bien alguno no consiga maximizar su valor dadas las decisiones de los demás, o que algunos no tomen la decisión que sería mutuamente maximizadora para ellos, dadas las decisiones de los otros.

Un principio que prescribe una decisión para cada agente y luego, suponiendo que tal guía es obedecida por algunos, prescribe una decisión diferente a los demás, provoca algo más que consternación y perplejidad entre sus posibles seguidores. La inconsistencia de los principios egoístas nos obliga a pensar en ciertos problemas de la interacción,

a reevaluar cuál es el fallo, en vista de que ciertos conflictos son ineludibles. Aunque cualquiera pueda lograr un objetivo egoísta, el hecho de que no todos puedan nos obliga a reconsiderar dónde reside la responsabilidad por el fallo en las situaciones del tipo de las examinadas. Normalmente disculpamos, al menos en parte, a una persona que no logra su objetivo cuando reconocemos que hizo todo lo que estuvo en su mano, ¿no deberíamos también disculpar a quien no haga todo lo que está en su mano cuando llegamos a la conclusión de que no todos pueden hacerlo?

Dejo éstas, y otras implicaciones de nuestro análisis sobre la inconsistencia del egoísmo, a la reflexión del lector. Quizá pudiera hallarse un remedio parcial para el conflicto entre la maximización individual y la coordinación entre varios agentes —tal como hubo un cierto remedio para el fracaso de la elección egoísta entre acciones mediante el expediente de la elección entre estrategias—. No hay que sacar una conclusión inamovible de nuestro tratamiento de la consistencia del egoísmo; no debiéramos precipitarnos al valorar sus implicaciones respecto a ciertos problemas, como la responsabilidad. De todos modos, vamos a dejar estos problemas a un lado, en parte porque otro más acuciante nos aguarda. Porque, pese a que hay límites para lo que los egoístas pueden hacer y decidir, hay, con todo, muchas situaciones en las que pueden de hecho tomar su mejor decisión. Pero, ¿deben hacerlo? El aspirante a egoísta, que hasta ahora no ve razón alguna para cambiar su actitud, quizá tenga que pensárselo dos veces si es posible mostrar que esa actitud es contraproducente.

## II. *¿Qué debe hacer un egoísta?*

1. «La auténtica razón de ser de una moralidad es producir razones que se impongan a las razones del autointerés en aquellos casos en que la adhesión de cada uno a su propio interés sea perjudicial para todos.»[21] Como tesis sobre las moralidades reales, esta afirmación de Kurt Baier tal vez sea falsa (o, en el mejor de los casos, sólo parcialmente cierta), pero como tesis sobre la moralidad racional —la moralidad acepta-

21. Baier, Kurt, *The Moral Point of View: A Rational Basis of Ethics*, Ithaca (N.Y.), Cornell University Press, 1958, pág. 309.

ble para agentes racionales— creo que no puede ser más verdadera. Una moralidad racional es una restricción, o conjunto de restricciones, a la maximización de valor relativo-al-agente, tales que quienes aspiran a maximizar tal valor encontrarían racional adherirse a ellas y cumplirlas.

Pero, ¿cómo puede ser racional para agentes maximizadores restringir su actividad maximizadora —o, en los términos de nuestra investigación—, como puede ser racional para los egoístas restringir su egoísmo? Me propongo responder ahora a esta pregunta, aunque durante algunos años pensé que carecía de respuesta, y así lo defendí. Cuando reflexioné por primera vez sobre la teoría moral de Kurt Baier, intenté comprender el conflicto entre razones de autointerés y razones capaces de imponerse al autointerés, y escribí, y pronuncié, una conferencia en la que estos problemas se perdían en un laberinto de palabras. Tras escuchar tales palabras, Howard Sobel me llamó aparte y, dibujando rápidamente una matriz en una hoja de papel, dijo: «Mira, de lo que estás hablando es del Dilema del Prisionero». Y miré, y fue como si cayeran las escamas de mis ojos y recibiera la luz.[22]

Pero al principio, mi visión fue tenue. En el Dilema del Prisionero vi una clara representación del conflicto entre razones interesadas y razones morales o cooperativas, pero ninguna parecía predominante. Vi un conflicto entre dos concepciones de racionalidad —una individual y prudencial, la otra colectiva y moral—. Y llegué a decir que «el individuo que, para ser moral, necesita una razón que no sea ya una razón moral, no puede conseguirla ... Porque es algo más que simplemente paradójico suponer que reflexiones basadas en el beneficio puedan justificar por sí mismas un perjuicio real».[23] Me equivoqué. Y quiero ahora desmentir esa paradoja supuestamente genuina —quiero mostrar que puede haber y hay de hecho una razón no moral para ser moral, una razón que incluso el egoísta tiene que reconocer.

El egoísmo es contraproducente. El objetivo del egoísta es que le vaya lo mejor posible: maximizar el valor relativo-al-agente. Típicamente, pensamos en el egoísta como alguien cuyo valor (lo que valora)

22. Este incidente ocurrió en la Universidad de California, Los Ángeles, en noviembre de 1965.

23. Gauthier, D., «Morality and Advantage», *Philosophical Review* 76 (1967), págs. 460-475. Trad. cast.: «La moral y la ventaja», en Joseph Raz (comp.), *Razonamiento práctico*, México, FCE, 1986.

se identifica con su interés o beneficio, y así suponemos que su objetivo es la maximización de ese valor concreto: su propio beneficio. Sin embargo, el egoísta no logra tal objetivo —de hecho, lo logra en menor medida que otros que no son egoístas—. Esto provoca una reflexión sobre el objetivo maximizador, que lleva al egoísta a encontrar una razón para cambiar su actitud. Con ese cambio arranca las escamas egoístas de sus ojos y comienza a ver como un ser moral —lo cual incluye aceptar perjuicios reales—. Embarcado en la travesía de la decisión racional, el egoísta descubre que su destino es, contra todo pronóstico, la teoría moral.

Mostrar que el egoísmo es contraproducente no es, sin embargo, tarea fácil. Como veremos, no basta con mostrar que, al tratar de maximizar un valor relativo-al-agente, los egoístas no logran que les vaya *colectivamente* tan bien como podría irles. No basta con mostrar, en palabras de Baier, que «la adhesión de cada uno a su propio interés sea perjudicial para todos». Lo que hemos de mostrar más bien es que la adhesión de cada persona a su propio interés es perjudicial para ella misma; que nadie logra que le vaya tan bien *individualmente* como podría irle. Sólo un argumento dirigido al egoísta como individuo será lo bastante convincente como para demostrarle que *su* actitud es contraproducente. Aun así, el desarrollo del argumento puede comenzar desde la perspectiva de todos —a partir del fracaso de los egoístas, como grupo, en su intento de conseguir el mejor resultado posible— para mostrar luego cómo se relaciona esta perspectiva con la de cada individuo. Podemos comenzar, por tanto, con el Dilema del Prisionero.

2. Frente a la escasa atención prestada por los filósofos interesados en estudiar la consistencia del egoísmo a los problemas que plantea el equilibrio de Nash, el Dilema del Prisionero se ha convertido ya en un tema habitual, debido a la evidencia con que muestra el carácter aparentemente contraproducente del comportamiento egoísta. Recordemos qué es exactamente lo que muestra el dilema. Cada prisionero —Jack y Zack, como les llamé en la primera parte— dispone de una estrategia que es la mejor respuesta a cualquier estrategia que el otro elija. Esta estrategia es confesar. Pero el resultado si cada uno elige su mejor respuesta, es desventajoso para ambos. A ambos les habría ido mejor si hubieran elegido la estrategia alternativa: no confesar, mantener su silencio. El mejor resultado que cada uno puede alcanzar,

haga el otro lo que haga, lo logra confesando; pero cada uno obtiene un resultado mejor si ninguno confiesa que si ambos lo hacen.

Necesitamos un término más para hablar con precisión sobre el dilema: el «óptimo de Pareto».[24] Un resultado es óptimo de Pareto sólo si no hay ninguna alternativa factible que procure mayor valor para una persona sin dejar a alguna otra con un valor inferior. O, suponiendo que el valor y la preferencia se corresponden de alguna manera, y sólo si no hay ninguna alternativa que sería preferida por alguien sin ser inadmisible para nadie. También se puede formular diciendo que un resultado es óptimo de Pareto sólo si todas las alternativas factibles que conceden a alguna persona mayor valor, implican que otra persona reciba menor valor.

Pensemos en los resultados posibles para Jack y Zack. Si ambos confiesan, cada uno recibe una dura sentencia, pero que no llega al máximo. Si ninguno confiesa, cada uno recibe una sentencia leve, pero que excede el mínimo. Si uno confiesa y el otro no, quien confiese recibe la sentencia mínima, y el otro la máxima. Supongamos que sus valores estén en función inversamente proporcional a la duración de la condena. Ahora, si comparamos cada posible par de resultados, veremos que en todos los casos Jack prefiere un miembro del par y Zack el otro, excepto que ambos preferirán el resultado de que ninguno confiese al de que ambos lo hagan. Por tanto, el resultado de la confesión de ambos *no* es óptimo de Pareto; hay una alternativa que aporta a ambos mayor valor. Todos los demás resultados *son* óptimos de Pareto; para todos ellos ocurre que las alternativas que aportan más valor a uno de los prisioneros, implican una disminución de valor para el otro.

Pero las estrategias que llevan a ambos a confesar están en equilibrio; cada una es la mejor respuesta a la otra. Y, puesto que se trata de la mejor respuesta a cualquiera de las posibles elecciones del compañero, no hay ningún otro conjunto de estrategias que constituya un equilibrio de Nash. Así pues, los resultados se dividen en dos conjuntos exclusivos y exhaustivos: el conjunto de los resultados que son óptimos de Pareto, y el conjunto de los resultados derivados de estrategias en equilibrio de Nash. Que los conjuntos sean exhaustivos no es una ca-

24. El término «óptimo de Pareto» hace alusión a Vilfredo Pareto, quien no habló de optimidad en absoluto, sino de «ofelimidad».

racterística común de las estructuras de interacción. Pero que sean exclusivos es normal, o al menos no es infrecuente, y es lo que convierte a este juego en un presunto dilema. Ello se debe a que si, como hemos argumentado, los egoístas bien informados se limitan a asegurarse los resultados derivados de estrategias en equilibrio, ello les impediría, en situaciones como la del Dilema del Prisionero, alcanzar un resultado óptimo. Cada uno consigue que le vaya lo mejor posible, dadas las circunstancias, pero a todos les podría ir mejor.

Un resultado puede concebirse de dos modos distintos, ambos importantes para la racionalidad. Por un lado, puede concebirse como el producto de los elementos de un conjunto de estrategias; por otro, puede verse como un conjunto de pagos. Concebido como el producto de un conjunto de estrategias, decimos que el resultado es un equilibrio de Nash y sólo si cada estrategia maximiza el valor del agente dadas las restantes estrategias. Considerado como un conjunto de pagos, decimos que un resultado es óptimo de Pareto sólo si cada pago maximiza el valor de quien lo recibe dados los restantes pagos.[25] No existe ningún principio completo de decisión en interacción, encargado de determinar un conjunto (no vacío) de decisión para cada conjunto de estrategias alternativas, tal que incluya en su conjunto de decisión algún miembro del conjunto de estrategias en equilibrio de Nash, y produzca además un resultado óptimo de Pareto. Ningún principio completo puede asegurar a la vez, en todas las interacciones, el equilibrio de Nash y la optimidad de Pareto. Éste es el teorema de imposibilidad, ilustrado por el Dilema del Prisionero, al que se enfrentan los egoístas y todos los maximizadores de valor relativo-al-agente.

Hay que advertir que esta imposibilidad no demuestra una inconsistencia adicional del egoísmo. Este problema no es como los discutidos en la primera parte. No hay ninguna dificultad para formular un principio de decisión egoísta en el Dilema del Prisionero u otras situaciones semejantes de incompatibilidad entre equilibrio y optimidad. La demanda de equilibrio se satisface directamente, sin necesidad siquiera de recurrir a seleccionar estrategias en vez de acciones. Y la condición del equilibrio es suficiente para determinar un principio egoísta;

25. Esto es verdad sólo si las funciones de pagos son continuas. La afirmación genérica es que un resultado es óptimo de Pareto sólo si cada pago maximiza el valor de quien lo recibe suponiendo que ningún otro pago sea reducido.

pues el problema de la coordinación es irrelevante en el Dilema del Prisionero. El requisito sobre la coordinación que introduje en la primera parte se aplica en las situaciones que tienen más de un conjunto de estrategias en equilibrio, cosa que no ocurre en el Dilema del Prisionero, donde sólo hay uno. Como maximizadores del valor relativo-al-agente, los egoístas sólo pueden coordinar sus estrategias dentro de los límites permitidos por el requisito de que cada agente considere su estrategia una mejor respuesta a las estrategias que espera que elijan los demás. Si cada agente tiene sólo una mejor respuesta, elija lo que elija el otro, tal como sucede en el Dilema, no hay lugar para la coordinación.

Para aclarar la diferencia entre un simple problema de coordinación y el Dilema del Prisionero, podemos comparar dos juegos. Revisemos en primer lugar el juego de las monedas entre dos personas que sirvió, en la sección 7 de la primera parte, para explicar el requisito de la coordinación. Aquí el juego tiene los siguientes valores monetarios:

| *Acción* | | *Pago* | |
|---|---|---|---|
| A | B | A | B |
| Cara | Cara | $2 | $2 |
| Cara | Cruz | –$2 | –$2 |
| Cruz | Cara | –$2 | –$2 |
| Cruz | Cruz | $1 | $1 |

En este caso, la mejor jugada de cada uno es mostrar lo mismo que el otro; esto asegura el equilibrio. A ambos les va mejor si eligen cara que si eligen cruz; de este modo, la coordinación mutuamente beneficiosa les permite elegir entre conjuntos de estrategias en equilibrio. Por el contrario, veamos ahora el siguiente juego, del tipo del Dilema del Prisionero:

| *Acción* | | *Pago* | | | |
|---|---|---|---|---|---|
| A | B | A | B | | |
| Cara | Cara | $2 | $2 | (1) | (2,2) |
| Cara | Cruz | –$3 | $3 | (3) | (4,1) |
| Cruz | Cara | $3 | –$3 | (3) | (1,4) |
| Cruz | Cruz | $1 | $1 | (2) | (3,3) |

Aquí también les va mejor a ambos si juegan cara que si juegan cruz. Pero un principio de decisión egoísta nunca tendrá esto en cuenta. Porque la mejor estrategia para ambos es elegir cruz, haga lo que haga el otro; sólo ese producto de estrategias asegura el equilibrio. Y así, el requisito del equilibrio no deja lugar para otras consideraciones. Los egoístas *no* deben jugar cara, porque no es lo mejor que pueden hacer (ninguno de ellos). Quizá lamenten el resultado de su comportamiento egoísta pero, al elegir cruz, cada uno actúa lo mejor que puede para favorecer sus intereses. La pregunta es ¿debe hacerlo? ¿debe *ser* egoísta?

Antes de intentar responder esta cuestión, hemos de comprender el conflicto general entre equilibrio y optimidad que la concreta estructura del Dilema del Prisionero ejemplifica. Conviene no malinterpretar la naturaleza de este conflicto. Un egoísta, cuya preocupación es la maximización de valor relativo-al-agente, es del todo indiferente tanto al equilibrio como a la optimidad. Lo único que le importa es su propio pago. Si las estrategias de los demás ya están dadas, elige la estrategia más beneficiosa para sí mismo; si todos se comportan así, el equilibrio resulta de modo involuntario. Si los pagos de los demás estuvieran dados, y él tuviera que elegir entre pagos posibles, elegiría el más beneficioso para sí mismo. Si todos se comportaran según este patrón, el resultado sería un óptimo de Pareto, también de modo involuntario. Lo que al egoísta le interesa son los resultados, pero como el ámbito de su poder de decisión son las acciones o estrategias, su interés por los resultados puede expresarse únicamente en su elección de estrategias. Lo que el Dilema del Prisionero demuestra es que en algunas situaciones, su decisión no expresa de modo efectivo su interés. Al elegir entre estrategias, el egoísta puede ser incapaz de maximizar su pago dados los pagos de otros; incapaz, por tanto, de obtener un beneficio del que podría disfrutar sin coste para los demás. Podemos decir, en conclusión, que el egoísta se enfrenta a un *conflicto entre estrategia y pago*. Éste es el problema general que el Dilema del Prisionero revela.

3. ¿Hasta qué punto es importante el conflicto entre estrategia y pago? Lo es desde luego en el sentido de que, cuando ocurre, complica la vida a los egoístas (y, en general, a todos los maximizadores de valor relativo-al-agente). Pero, ¿ocurre alguna vez fuera de las estructuras de interacción analizadas por los teóricos de juegos? Por-

que si resulta ser un fenómeno carente de relevancia práctica, difícilmente podrá servir como base para argumentar que el egoísmo es contraproducente.

Sin embargo, sí ocurre. De hecho, el conflicto entre estrategia y pago es uno de los fenómenos fundamentales en la vida social. Constituye la médula del problema de asegurar la oferta óptima de bienes públicos o colectivos. Explica la suboptimidad típicamente resultante de no internalizar ciertos efectos —la presencia de un coste social neto de la contaminación mayor que el beneficio individual neto de quienes contaminan—. Es el fundamento de la «tragedia del terreno comunal», de Garret Hardin,[26] y ayuda a explicar la observación de John Kenneth Galbraith cuando afirma que una sociedad rica disfruta de «opulencia privada y miseria pública».[27] Nos permite comprender por qué incluso un gobierno que gastara sabiamente sus recursos, necesitaría una agencia tributaria para recaudarlos. El contexto del conflicto entre estrategia y pago es campo abonado para «gorrones» (*Free-riders*) y parásitos.

La importancia de este conflicto, hoy ampliamente reconocida, fue durante mucho tiempo ignorada por nuestro pensamiento social y económico. En primer lugar, los economistas tienen, desde Adam Smith, una injustificada tendencia a centrarse en el mercado perfectamente competitivo —en cuya feliz estructura no hay el menor conflicto entre estrategia y pago—. Como elocuentemente ha dicho Rusell Hardin, el Dilema del Prisionero es el dorso de la «mano invisible».[28] En el mercado perfecto, la mano invisible asegura que si cada uno intenta promover su interés particular, el interés social es fomentado, aunque de modo involuntario. Dicho más precisamente, la actividad del mercado —en la que cada individuo trata de maximizar el valor de una función definida sobre los bienes que consume y los factores de producción (servicios) que ofrece— lleva a un resultado que está (en términos de utilidad) sobre la frontera de posibilida-

26. Véase Hardin, G., «The Tragedy of the Commons», *Science*, 162 (1968), págs. 1.243-1.248.

27. Véase Galbraith, J.K., *The Affluent Society*, Boston, Houghton Mifflin, 1958, pág. 257. Trad. cast.: *La sociedad opulenta*, Barcelona, Ariel, 1992 (tercera ed.). Galbraith no emplea, sin embargo, este tipo de explicación.

28. Véase Hardin, R., *Collective Action*. Baltimore, John Hopkins University Press, 1982; pág. 7.

des de la sociedad, de modo que no podría mejorarse la situación de una persona sin empeorar la de otra. El equilibrio resultante de los intercambios voluntarios es un óptimo de Pareto.

Si el mundo fuera un mercado perfectamente competitivo —tal como sostiene la doctrina imputada a algunos economistas de la escuela de Chicago— los egoístas no tendrían razón alguna para cambiar su actitud directamente maximizadora. El Dilema del Prisionero sería una curiosidad lógica, reveladora de un tipo de interacciones, afortunadamente irreales, en las que los egoístas no lograrían alcanzar la frontera de posibilidades de utilidad y, por tanto, no lograrían que les fuera, colectivamente, tan bien como podría irles. Pero a pesar de lo sugerente que resulta el mercado, que nos muestra la posibilidad de interacciones que no dan lugar a ningún problema entre maximizadores de valor relativo-al-agente —nos muestra, de hecho, un tipo de interacción que no necesitaría la guía de esos principios limitadores del comportamiento maximizador que constituyen una moral racional—, no puede decirse que el mundo real, según lo vemos la mayoría de nosotros, se asemeje demasiado al reino de la competencia perfecta. Así, estamos abocados a enfrentarnos a conflictos entre estrategia y pago, tanto en nuestras interacciones cotidianas, como a la hora de diseñar las instituciones sociales que estructuran esas interacciones.

Aparte del mercado perfecto, hay un segundo factor que ha dificultado la comprensión de este conflicto. De hecho, la conciencia del fracaso del mercado como modelo de una gran parte de la interacción social no implica el correcto reconocimiento del problema fundamental que ha de enfrentar la acción pública o colectiva que no es de mercado. La tentación inmediata es suponer que, al igual que un individuo racional decidirá lo que, dada la información disponible, es mejor para él, un grupo de individuos racionales decidirán lo mejor para ellos, dentro de lo posible. Extrapolamos de la acción individual a la acción colectiva o de grupo.

Sin embargo, en esta extrapolación se esconde, por lo general, una falacia, cuyo reconocimiento debemos a Mancur Olson Jr., quien al parecer fue el primero en referirse a ella, en su libro *La lógica de la acción colectiva*.[29] Aclararé en qué consiste la falacia con un ejemplo

29. Olson, M., *The Logic of Collective Action*, Cambridge (Mass.), Harvard University Press, 1965. Trad. cast.: *La lógica de la acción colectiva: Bienes públicos y la*

tomado, no del libro de Olson, sino del de Russell Hardin.[30] Supongamos que diez unidades de un bien público puro, que sólo puede ofertarse en común, están disponibles para una comunidad de diez personas. Cada unidad cuesta $5, y proporciona a cada miembro de la comunidad un beneficio de $1. Cada uno debe decidir si contribuir con $5 para la provisión social del bien. Si todos contribuyen, el beneficio total es $100, y el coste total $50, lo que produce un beneficio social neto de $50 y un beneficio neto para cada individuo de $5. Si nadie contribuye, el beneficio neto, social e individual es $0. Sin embargo, nadie que pretenda maximizar su pago, contribuirá. Cada uno supondrá que n personas (entre 0 y 9) contribuirán. Si él también contribuye, se producirá un bien más y el beneficio social neto sería $5(n+1), que se dividirá de modo que el beneficio neto para cada uno de los que contribuye es $(n-4), y para cada uno de los que no contribuye $(n+1). Si no contribuye, el beneficio social neto es $5n, el cual se divide igualmente y arroja un beneficio neto para cada miembro contribuyente de $(n-5), y para cada miembro no contribuyente, de $n.[31] Así, si contribuye, su beneficio neto es $(n-4); si no contribuye, su beneficio neto es $n. Y para todos los valores posibles de n, $n es mayor que $(n-4), de modo que decide no contribuir.

Es evidente la analogía entre este razonamiento y el del Dilema del Prisionero. Cuando se trata de un bien público puro, cada individuo decide no contribuir, y aprovecharse del bien; el resultado es, lógicamente, que no hay nada de lo que aprovecharse. El caso analizado es artificialmente simple; en casos más reales, en los que el

---

*teoría de grupos*, México, LIMUSA, 1992; la falacia está expuesta en la introducción, págs. 1-2.

30. Hardin, R., *op. cit.*, págs. 25-28.

31. Por ejemplo, supongamos que el agente espera que cinco personas contribuyan (n=5), de modo que se producirán cinco unidades del bien. Si él contribuye, se producirán seis unidades. Como cada persona obtiene un beneficio de $1 por unidad, el beneficio individual será de $6. Quienes contribuyen habrán pagado $5, con lo que obtienen un beneficio neto de $1 [$(n-4)], mientras que, para quienes no contribuyen el beneficio neto es $6 [$(n+1)].

Si, por el contrario, no contribuye, se producirán un total de cinco unidades. Como cada persona obtiene un beneficio de $1 por unidad, el beneficio individual será $5. Quienes contribuyen habrán pagado $5, con lo que su beneficio neto será $0 [$(n-5)], mientras que, para quienes no contribuyen, el beneficio neto es $5 [$n].

coste de cada unidad disminuye al aumentar su número, puede ser que para algunos individuos merezca la pena soportar el coste de pagar para que haya algunas unidades del bien para todos, pero antes de llegar al nivel óptimo de oferta, cada cual preferirá aprovecharse del nivel alcanzado, sin contribuir para la producción de una unidad adicional.

El reconocimiento del problema de la acción colectiva disipará sin duda cualquier tentación de suponer que mi argumento no se dirige a nosotros, sino a una clase diferente de personas —los egoístas—. Todos nosotros somos maximizadores de valor relativo-al-agente —o algo muy cercano a ello— en muchas de nuestras interacciones. Y todos nos enfrentamos al problema de la acción colectiva planteado por el dorso de la Mano Invisible. En esa situación nuestro comportamiento tiende a ser, como el del egoísta, contraproducente. Eso es al menos, lo que yo defiendo. Ahora debo probarlo.

4. No parece problemático aceptar que el egoísmo es contraproducente en las situaciones en que se da un conflicto entre estrategia y pago. El resultado de la interacción egoísta concede a cada agente un pago inferior, cuando podría obtener otro mayor sin disminuir en absoluto los pagos de los demás. Todos ganarían en ese caso; luego hay un beneficio neto posible que no se produce. El resultado es que les podría ir mejor de lo que les va y, por tanto, el comportamiento egoísta frustra su propio fin. El egoísta intenta maximizar su valor, y acaba obteniendo menos que el no egoísta que intenta, en cambio, alcanzar un óptimo de Pareto.

Sin embargo, no deberíamos dejarnos convencer por este argumento. *Ninguno consigue* tanto como podría, y a *todos les va* peor de lo que podría irles. De ahí no se sigue que a *cada uno le vaya* peor de lo necesario. Para mostrar que el egoísmo es contraproducente, no basta con pensar en su resultado global, sino que hay que analizar la situación del individuo que ha de decidir en función de las decisiones esperadas de los demás. Juzgar la situación que afrontan los egoístas exclusivamente desde el punto de vista global o colectivo, supone cometer otra versión de la falacia denunciada por Olson en su análisis de la acción colectiva. Se pensó, ingenuamente, que un grupo de individuos maximiza su beneficio neto global del mismo modo en que lo hace un solo individuo. Hemos reconocido la falacia de ese postulado, pero ahora estamos suponiendo que un individuo fracasa en su

intento de maximizar su beneficio neto del mismo modo que le sucede a un grupo. Sin embargo, al igual que la maximización desde el punto de vista individual no tiene por qué coincidir con la maximización desde el punto de vista de un grupo, lo que es contraproducente desde el punto de vista del grupo no lo es necesariamente desde el punto de vista de un miembro individual. Y sólo desde este último punto de vista podemos convencer al egoísta de que su actitud es contraproducente.

El egoísta nos dice que aún no le hemos demostrado nada: él elige su mejor respuesta, dadas sus expectativas sobre lo que harán los demás, ¿qué más puede hacer? El problema —si es que se trata de un problema y no de simple mala suerte— es que las decisiones de otros dan lugar a que él consiga menos de lo que podría, dados los pagos de todos. Es víctima de las decisiones de los demás, no de las suyas propias. Pero tampoco es que ellos traten de perjudicarle; lo único que hacen es decidir en vista de sus intereses particulares. Cada uno acaba siendo víctima de decisiones que no tienen en cuenta sus costes y beneficios personales. Pero ser víctima de una situación no es lo mismo que actuar de modo contraproducente. De hecho, si el egoísta se quejara de que los demás no tienen en cuenta sus costes y beneficios, ellos replicarían, con toda justicia, que si hubieran querido tener en cuenta algo más que sus propios intereses, entonces sí se les podría achacar un comportamiento contraproducente.

La existencia del conflicto entre estrategia y pago perjudica a los egoístas. Pero ningún egoísta en concreto sufre tal menoscabo, ya que ninguno podría mejorar su resultado. Decidir de otro modo —con una actitud no egoísta— no remediaría el problema; el único remedio es evitar que surjan los conflictos entre estrategia y pago. Quienes esperan correr con los costes de tales conflictos tienen, así, una buena razón para establecer sanciones —mediante acuerdos obligatorios o algún sistema coactivo externo— que alteren los pagos, de modo que las estrategias en equilibrio den lugar a resultados óptimos de Pareto. Éstos son los métodos clásicos propuestos por Thomas Hobbes, mucho antes de que los teóricos de juegos explicitaran la estructura precisa que da lugar al conflicto. Los pactos —pero no pactos sin la espada, porque ésos quedan en meras palabras, que carecen de la fuerza necesaria para ofrecer garantías a un hombre— y el soberano, que impone los pactos y estructura las instituciones sociales de tal modo que pre-

vengan los actos de los «gorrones», traen orden al mundo de los egoístas.[32] Estos métodos cautelares implican costes que los egoístas preferirían evitar —un aspecto al que quizás Hobbes no concedió suficiente importancia—,[33] pero el mundo no tiene ninguna obligación de acomodarse a todas nuestras preferencias. No hay nada contraproducente en la necesidad de vérnoslas con estructuras de interacción que dificultan la acción de quienes tratan de realizar sus valores relativos-al-agente en la mayor medida posible.

Parece que la acusación de que el egoísmo es contraproducente descansa en una confusión. Hay que distinguir entre las decisiones de los individuos y la estructura dentro de la cual las toman. La afirmación de que el egoísmo es contraproducente se refiere a los efectos de las decisiones egoístas, que no puede ser defendida simplemente mostrando los efectos del conflicto entre estrategia y pago. Estos efectos limitan las posibilidades de decisión; pero dentro de esos límites, los egoístas actúan del mejor modo posible. Si hubiera otras posibilidades (si el mundo fuera un mercado perfecto) les iría mejor.

Un egoísta inteligente intentará, desde luego, convencer a los demás de que el egoísmo es una política contraproducente. Consciente de los costes que los demás imponen sobre él, intentará, en su propio interés, manipularlos para que no lo hagan. Quizás apele a la idea del conflicto entre estrategia y pago con la esperanza de convencerlos de que, dado que todos obtienen menos de lo que podrían, su egoísmo es contraproducente. Mas se trata de una argumentación especiosa, cuyo propósito es garantizarse los beneficios del comportamiento no egoísta de los demás, mientras él, astutamente, sigue desplazando sobre ellos tantos costes como pueda —con la única condición de fingir que hace lo que predica para que su prédica sea efectiva—. Así pues, la tesis de que el egoísmo es contraproducente quizá no sea tanto el producto de una inadecuada comprensión del conflicto entre estrategia y pago, como una deliberada distorsión de su carácter real, maquinada por el propio egoísta.

32. Hobbes, T., *Leviathan*, Londres, 1651 (ed. de C.B. Macpherson, Harmondsworth, Middlesex, Penguin, 1968), caps. 15 y 17.

33. Aunque sí le dedicó cierta atención. Al final del capítulo 18 del *Leviatán*, advierte que «ningún Estado puede existir sin una u otra incomodidad», para continuar comparando los costes del gobierno con los de la guerra civil y la carencia de toda autoridad.

5. Mas este alegato del egoísta es equivocado. En definitiva, el egoísta no obtiene todo lo que podría. No dudo que, ante esta afirmación, y en vista de lo dicho hasta ahora, el lector estará en guardia; puede que yo sea el egoísta, tratando de engañarle con mis melifluas palabras. Contra esta sospecha, lo único que puedo ofrecer es mi argumento, el cual se cifra en lo siguiente: indudablemente, un egoísta maximizará el valor relativo-al-agente siempre que pueda. Estaremos de acuerdo en que, fuera de las situaciones en que el egoísmo no ofrece una guía consistente para la decisión, el egoísta elige siempre una mejor respuesta a las decisiones esperadas de los demás y que, *en esas situaciones*, eso es lo mejor que puede hacer. Ahora bien, su modo de decidir condiciona el tipo de situaciones en que puede esperar encontrarse. Y el efecto de ese condicionamiento es desventajoso para él. El egoísta aprovecha al máximo sus oportunidades; pero, precisamente por ser egoísta, tendrá menos oportunidades que los no egoístas —evidentemente, esto no se refiere a cualquier tipo de no egoísta, sino a un tipo especial, que denominaré «cooperador»—. Éste es el argumento cuyo desarrollo nos hará ver el carácter contraproducente del egoísmo.

En el Dilema del Prisionero se puede distinguir una estrategia cooperativa y otra no cooperativa para cada agente. La estrategia no cooperativa es, por supuesto, confesar; la cooperativa es callar. Sería de agradecer que los prisioneros resultaran ser individuos no cooperadores, que no hubiera honor entre los ladrones, pero en general —y desde el punto de vista de los afectados, siempre— la no cooperación es costosa. A dos cooperadores les iría mejor (a cada uno de ellos) que a dos no cooperadores. El problema es, como hemos visto, que un no cooperador emparejado con un cooperador obtendría un resultado aún mejor, a expensas del cooperador.

Pero supongamos que un agente estuviera *condicionalmente* dispuesto a cooperar en las situaciones del tipo del Dilema del Prisionero y, en general, en todas las situaciones en que se produce el conflicto entre estrategia y pago. No es que fuera a optar por una estrategia cooperativa sin más. Formaría una expectativa sobre las decisiones estratégicas de los demás compañeros (u oponentes) y basaría su propia decisión en esa expectativa. Elegiría la cooperación como respuesta a la cooperación esperada y la no cooperación como respuesta a la no cooperación esperada.

¿Cómo afectaría esta disposición cooperativa condicional a sus pagos? A primera vista, parece que los reduciría. Cuando esperara que el otro eligiese una estrategia no cooperativa, su decisión maximizaría su pago esperado. Pero cuando esperara que el otro agente cooperase, no maximizaría su pago esperado, de modo que ganaría menos que si eligiera siempre no cooperar.

Pero este argumento no tiene en cuenta la disposición del otro agente o agentes. Supongamos que el otro agente también estuviera condicionalmente dispuesto a cooperar. Entonces, si nuestro agente no estuviera dispuesto a cooperar, y confiase en que el otro podría reconocer su intención, no esperaría la cooperación de su parte, con lo que acabarían ambos en el desventajoso resultado producto de las estrategias en equilibrio. Por contra, si estuviera dispuesto a cooperar, y confiara en que el otro lo advirtiese, esperaría la cooperación de su parte y, por tanto, que el resultado fuera un óptimo de Pareto, beneficioso para ambos. Entre cooperadores condicionales, se produce tal relación de las expectativas sobre las decisiones de los demás con las propias disposiciones a decidir, que todos pueden beneficiarse de la interacción de un modo que los no cooperadores no pueden emular.

El egoísta trata de maximizar el valor relativo-al-agente dadas sus expectativas sobre las estrategias de otros. Pero las decisiones de los demás y, por tanto, las expectativas sobre las que el egoísta funda la suya, pueden verse condicionadas por su política egoísta y maximizadora en el sentido de que otros, anticipando su decisión, responden con actitud maximizadora. El cooperador se abstiene de perseguir la maximización del valor dadas sus expectativas sobre las estrategias de otros. Puede así influir en las decisiones ajenas y, por ende, en sus expectativas propias: otros cooperadores, anticipando su decisión, responden con una actitud cooperativa. Por eso el egoísmo es contraproducente. Nuestro argumento se apoya en una comparación entre los efectos de tomar como base de la decisión la maximización no cooperativa y los efectos de adoptar la cooperación condicional. Aunque el cooperador condicional se abstiene de sacar el máximo partido a sus oportunidades, acaba por encontrar oportunidades de las que el egoísta carece y, por tanto, puede esperar que sus pagos sean superiores a los de éste.

Evidentemente, el cooperador condicional puede errar. Puede que no reconozca la disposición cooperativa de otros hacia él, y les trate como egoístas; puede no reconocer el egoísmo de otros y, al tratar-

les como cooperadores, dar lugar a que se aprovechen de él. A menos que los cooperadores sean razonablemente capaces de identificarse mutuamente y de distinguir a los no cooperadores, su disposición condicional puede resultar desventajosa. Ésta es una cuestión empírica. Sin embargo, dados los beneficios reales de la cooperación, es de esperar que los cooperadores condicionales potenciales traten de desarrollar su pericia tanto para identificar las disposiciones de aquellos con quienes interactúan, como para dar a conocer su propia disposición. Lo beneficiosa que realmente sea la cooperación condicional dependerá de esta pericia, así como de la proporción de cooperadores sobre el total de agentes en la interacción. Mas el beneficio potencial de la disposición no depende de factores empíricos; es reflejo de la estructura lógica de la interacción. Idealmente, un individuo cuyo objetivo es egoísta (que le vaya lo mejor posible), debe esperar que le vaya mejor como cooperador que como egoísta.

Sostengo, pues, que, dada la capacidad de elegir entre el egoísmo y la cooperación condicional, y dada también la pericia suficiente para identificar las disposiciones de otro y facilitar la identificación de las propias, una persona racional decidirá disponerse a cooperar condicionalmente. Esta decisión es ella misma egoísta (supuesta una elección entre disposiciones, la persona maximiza su valor relativo-al agente esperado al optar, en concreto, por la cooperación condicional). Pero su efecto es transformar a una persona egoísta en cooperadora, en alguien que, bajo determinadas circunstancias, no elige de modo egoísta.

Antes de analizar las objeciones a este argumento, debo advertir que *no* depende de la suposición de que, como los agentes se encontrarán probablemente con una serie indefinida de conflictos entre estrategia y pago, elegir la cooperación en cada situación concreta afectará a las expectativas (y, por tanto, a las decisiones) de otros en la ocasiones posteriores. Se ha prestado considerable atención a la importancia de la reputación en los Dilemas del Prisionero repetidos, así como en situaciones en las que el agente se beneficia de la disuasión proporcionada por una amenaza creíble —por ejemplo, la situación de entrada en el mercado analizada por Reinhard Selten en «la paradoja de la cadena comercial»—.[34] Pero nuestro objetivo no son los con-

34. Véase, Selten, R., «The Chain-store Paradox», *Theory and Decision*, 9 (1978), págs. 127-159.

ceptos de reputación o amenaza. El fundamento para preferir la cooperación condicional al egoísmo no depende de la suposición de que se pueden obtener beneficios a largo plazo gracias a la reputación adquirida a base de decisiones cooperativas. Mi argumento es aplicable a un conflicto que ocurra una sola vez.[35]

Supongamos que cada persona supiera —no importa cómo— que una (y sólo una) vez en su vida, ha de afrontar un conflicto entre estrategia y pago. Si esta persona pudiera identificar con certeza la disposición de los demás agentes en esa situación, y esperara que los demás identificaran igualmente la suya, estaría racionalmente justificado para ella adoptar la disposición condicionalmente cooperativa. Porque si lo hiciera, y los demás también resultaran ser cooperadores condicionales, lograría que le fuese mejor que con una actitud no cooperativa. Y si los demás no fuesen cooperadores, no le iría peor que con la actitud egoísta (puesto que les habría identificado correctamente y actuado en consecuencia). Es interesante advertir que en esta situación, una persona que se encontrase entre cooperadores *no* maximizaría el valor relativo-al-agente, ni en términos de sus expectativas a corto plazo (referidas a la situación concreta), ni en términos de expectativas a largo plazo (para lo que le quedara de vida). Si está genuinamente dispuesta a cooperar, no se comportaría en modo alguno como un egoísta.

Pasemos ahora a las objeciones. Se dirá, sin duda, que aunque quizá sea racional fingir que se es un cooperador, no es racional serlo de verdad. Por mucho que lo parezca, el egoísta racional no habrá abandonado en realidad su egoísmo. No niego que pueden existir circunstancias en que la impostura sea la política racional para un maximizador de valor relativo-al-agente. Pero ni se ha propuesto, ni puede proponerse, un argumento que demuestre que éste es siempre el caso. Cabe, por ejemplo, que la impostura no funcione —si las capacidades detectivescas de los demás son demasiado buenas—. Puede ser, por otro lado, que la tensión psicológica de vivir en la impostura sea insoportable. Quizás el mejor modo de cosechar las ventajas de la cooperación consista en ser un cooperador genuino. La honestidad de la que habitualmente se dice que es la mejor política, puede resultar ser aquella que, una vez adoptada, ya no puede desecharse.

35. La aplicación a la disuasión puede verse en mi «Deterrence, Maximization and Rationality». *Ethics*, 94 (1984), págs. 474-495.

Puede decirse, por tanto, que nuestro argumento muestra que el egoísta debe reconocer los beneficios que en ciertas circunstancias concretas proporciona la adopción de la cooperación condicional. Pero entonces —se objetará en segundo lugar— el egoísmo no es contraproducente. Al contrario: contiene recursos para su propia reforma; y así resulta que la cooperación condicional es el verdadero egoísmo. Esta objeción interpreta el egoísmo de un modo muy distinto a como yo lo he hecho. Tanto el análisis sobre la «contraproductividad» del egoísmo, como el estudio de su posible inconsistencia, se referían a un principio egoísta de decisión entre estrategias o acciones. He mostrado que es contraproducente estar incondicionalmente dispuesto a actuar de acuerdo a tal principio —esto es, de acuerdo a un principio cuya condición consiste en determinar, para cada conjunto de estrategias alternativas posibles, un subconjunto cuyos elementos maximizan cierta medida relativa al agente definida sobre el conjunto inicial—. Interpreto el egoísmo como la disposición incondicional a actuar de acuerdo a tal principio. La persona que, por la razón que sea, decide no actuar conforme a ese principio, decide no ser un egoísta. El fundamento de esa decisión puede ser egoísta, lo mismo que la decisión misma. Mas lo decisivo es lo que se elige, no por qué, o cómo, se ha elegido; no se es egoísta si no se opta por un principio egoísta de decisión.

Aunque, ciertamente, el egoísmo contiene los recursos para su propia reforma. El egoísta es capaz de reconocer el carácter contraproducente de su actitud ante las decisiones y, así, posee una razón para seleccionar una actitud diferente. Pero la reforma que el egoísta lleva a cabo no es una modificación interna que deje invariable su actitud egoísta original. Elegir la cooperación condicional es el último acto del egoísta como egoísta, y en tal acto se afirma el carácter contraproducente del egoísmo.

6. Con esto no sólo limitamos las pretensiones del egoísmo, sino que también aumentamos las perspectivas para la moralidad. Si los principios de decisión egoístas no fueran contraproducentes, el moralista se vería abocado a rechazar, bien la concepción relativista del valor, bien la concepción maximizadora de la racionalidad. La moralidad, tal como la entiendo aquí, ofrece una restricción *interna* al intento directo de que a uno le vaya lo mejor posible. Una restricción interna es aquella que se interpone entre la evaluación del agente y su

decisión —una restricción, por tanto, sobre su principio de decisión, o función que determina, a partir de los conjuntos de estrategias alternativas, los conjuntos de decisión—. Tal restricción no podría ser legítima desde el punto de vista del egoísta. Por contra, una restricción *externa* sí sería explicable desde ese punto de vista ya que, para el egoísta, toda justificación lo es «para alguien», y quienes hubieran impuesto la restricción podrían muy bien encontrarla provechosa. Pero una restricción externa afecta, bien a las opciones existentes —y, en consecuencia, a las estrategias disponibles—, bien al valor que puede esperarse del resultado de algunas de las opciones. Una vez impuesta, una restricción externa deja al egoísta igual de libre para decidir conforme a su principio maximizador y, por tanto, deja su egoísmo intacto. Pero la adhesión voluntaria a una restricción sobre el comportamiento maximizador —una restricción que no influye en las opciones del agente ni en sus valores— es incompatible con el egoísmo. Si la moralidad representa tal restricción interna, entonces un principio de decisión moral debe ser incompatible con el egoísmo y, por tanto, una alternativa al mismo.

Para evitar posibles malentendidos, he de advertir que al tratar la moralidad como una restricción interna a la maximización directa, estoy ofreciendo una descripción necesaria, pero no suficiente, de la misma. No toda restricción interna concebible sería moral. No puedo detenerme aquí para analizar completamente qué habría que añadir a la idea de restricción interna para capturar el concepto de moralidad; lo que añadiré se refiere a la relación entre la moralidad y la cooperación.

Tradicionalmente, el rechazo del moralista hacia el egoísmo ha seguido una de las siguientes líneas de ataque (o las dos). La primera gira en torno a la concepción relativista del valor que defiende el egoísta. Como apunté en la primera parte, algunos filósofos, como G.E. Moore, creen que esta concepción es contradictoria, y sostienen que el valor debe ser absoluto. Acusan al egoísta de maximizar su interés, en vez de maximizar lo que es realmente bueno. Quizá su interés sea parte de ese «bien», pero no más que el interés de cualquier otra persona. Otros argumentan que el egoísta confunde su interés aparente con su verdadero interés, y defienden que el verdadero interés de cada persona está vinculado a un valor trascendente, no relativo, que permite que el conflicto reinante entre los intereses aparentes de los individuos sea

sustituido por la armonía de sus verdaderos intereses. Creo que ésta es la tesis que Platón propone en *La República*; los pensadores que profesan esta tesis acusan al egoísta de maximizar la realización del bien aparente, en vez del verdadero bien.

La segunda línea de ataque gira en torno a la concepción maximizadora de la racionalidad propia del egoísta. Algunos, como Kant, por ejemplo, argumentarían que el egoísta supone erróneamente que la razón es meramente instrumental —capaz sólo de determinar los mejores medios para fines dados— y no reconocen que la razón tiene un papel práctico independiente del establecido por el interés.[36] Desde este punto de vista, la moralidad surge al adscribir a la racionalidad práctica la misma universalidad hallada en la racionalidad teórica. Así como la razón teórica descubre leyes descriptivas o explicativas, la razón práctica descubre leyes prescriptivas o justificatorias. El compromiso del egoísta con la maximización es rechazado porque no alcanza la universalidad inherente a la verdadera racionalidad.

Ninguna de estas líneas de ataque contra el egoísmo me parece prometedora. Afortunadamente, no hace falta que defienda aquí esta afirmación; sería absurdo y presuntuoso pensar que puedo refutar dos de las principales tradiciones de pensamiento moral con unas cuantas palabras. En vez de eso, las evitaré. Concediendo al egoísta su concepción del valor relativo-al-agente y de la racionalidad maximizadora, he mostrado que si no abandona su egoísmo en favor de la cooperación condicional, se autoexcluye de oportunidades de interacción ventajosa con sus semejantes. Se puede esperar que la adhesión a un principio de decisión que establece las apropiadas restricciones sobre el comportamiento maximizador sea beneficiosa para quienes lo aceptan. Y así, partiendo de las premisas del egoísta pero rechazando sus conclusiones, encuentro un lugar para la restricción interna, para un principio de decisión no maximizador, moral.

Sugiero, por tanto, que mi ataque contra el egoísta requiere supuestos mucho más débiles que los ataques lanzados por los moralis-

36. Este párrafo pretende ser una interpretación de la tesis kantiana, tal como es desarrollada en la *Fundamentación de la metafísica de las costumbres*, y en la *Crítica de la razón práctica*, pero no pretendo defender los puntos controvertidos de mi interpretación. Sea o no fiel a Kant, me parece que mi explicación plantea importantes cuestiones sobre la instrumentalidad de la razón práctica, y sobre la conexión entre la razón práctica y la teorética; cuestiones que merecen una respuesta no kantiana.

tas tradicionales. Ellos asaltan su posición desde fuera, tratando de abatir sus fundamentos, mientras que yo la socavo desde dentro, mostrando que sus fundamentos no la sostienen. No necesito ni valor absoluto ni una racionalidad universalizada. Sugiero que los teóricos morales han recurrido a estas líneas de ataque porque no han visto la posibilidad de defender la moralidad combatiendo al egoísta en su propio terreno. Y sugiero, en fin, que el atractivo de las tradiciones platónica y kantiana dependía de no haber reconocido un tercer camino para que el moralista recupere el valor y la razón de las garras del egoísta.

Para añadir sustancia a estas sugerencias, veamos qué intuiciones positivas nos ofrece el fracaso del egoísmo sobre el lugar que puede ocupar la moralidad. El egoísmo no resuelve satisfactoriamente los conflictos entre estrategia y pago. Para resolverlos se necesita una disposición que he definido como condicionalmente cooperativa. La cooperación no es simple coordinación; el cooperador selecciona un curso de acción que promueve el beneficio mutuo y actúa en consecuencia, sin caer en la tentación de una defección que sería individualmente ventajosa. Así, la cooperación requiere alguna forma de restricción real. Si consideramos que la moralidad tiene que ver con la disposición a cooperar, entonces la moralidad será (o al menos incluirá) aquella parte de la teoría de la decisión racional que se ocupa de la formulación de principios para la interacción cooperativa. Estos principios cumplen el papel limitador tradicional de la moralidad de tal modo que su racionalidad debe ser reconocida por todos aquellos que, compartiendo los conceptos de valor y razón admitidos por el egoísta, hayan tomado conciencia del carácter contraproducente de sus decisiones.

En el Dilema del Prisionero la selección de las estrategias cooperativas no plantea problemas. Pero esto no es así siempre en los conflictos entre estrategia y pago. En el Dilema del Prisionero sólo hay un modo plausible de cooperar, porque sólo hay un resultado que es a la vez óptimo de Pareto y mutuamente beneficioso, en comparación con el resultado de equilibrio producto del comportamiento egoísta. Pero en la mayoría de los conflictos entre estrategia y pago hay muchos modos de cooperar —muchos resultados óptimos y mejores para todos los participantes que el resultado esperado si cada uno tratase de maximizar directamente el valor—. Los

principios morales deben sernos útiles para seleccionar alguna de estas posibilidades. Para poder ser usados eficazmente como alternativa al egoísmo, deben ser además razonablemente simples, y estar claramente establecidos en prácticas e instituciones sociales aceptadas. La cooperación depende de la capacidad de cada cooperador para anticipar las decisiones de los demás, y esto es posible con carácter general sólo si esas decisiones reflejan principios ampliamente compartidos.

Es previsible que los principios morales para la cooperación mutuamente beneficiosa requieran virtudes tradicionales como decir la verdad y mantener las promesas, como la honestidad, gratitud y benevolencia recíproca. Pero no hay por qué pensar que toda la moralidad tradicional va a superar el escrutinio del punto de vista cooperativo. La relación establecida entre la moralidad y la decisión racional pretende que la derivación de principios sea independiente de cualquier apelación a la práctica establecida. No estamos interesados en el equilibrio reflexivo.[37] Si ninguna restricción moral comúnmente reconocida pudiese relacionarse con la cooperación mutuamente beneficiosa —lo cual sería sorprendente—, la moralidad tradicional como tal no sería más que un cajón de sastre de ideas, sin un fundamento coherente. Mi teoría sobre la moralidad no pretende refinar nuestras ideas aceptadas, sino justificar sus restricciones sobre el firme fundamento de la decisión racional.

7. El papel de la teoría moral es ofrecer un criterio reflexivo y crítico mediante el cual puedan evaluarse y revisarse las prácticas morales existentes. El criterio para tal reflexión y crítica proviene del individuo que se pregunta cuál podría ser el sustituto racional de un principio egoísta —qué principio de decisión sería aceptable para él, si todos lo siguieran—. Debe pensar no sólo en sí mismo, sino en todos los demás: dado que su adhesión al principio va a quedar condicionada a sus expectativas sobre la adhesión de los demás, debe esperar que ellos también estén convencidos de su aceptabilidad.

Cada persona prefiere cooperar con los demás en los mejores términos posibles para sí misma. Pero cada uno debe reconocer que todos

37. Sobre la tesis de que la teoría moral ha de tener en cuenta el equilibrio reflexivo, véase Rawls, *A Theory of Justice*, Cambridge (Mass.) Harvard University Press, 1971, págs. 20-21, 48-51.

tienen esta preferencia. Por tanto, nadie esperará que otros —en la medida en que son racionales— acepten cooperar en términos menos ventajosos que los menos ventajosos que él mismo aceptaría. El reconocimiento de la racionalidad de todos conduce al requisito de que los principios morales sean mutuamente aceptables.

Esquemáticamente, podemos representar el problema de determinar un principio de decisión cooperativa mutuamente aceptable del siguiente modo. Llamemos C al conjunto de los principios factibles, y supongamos que el principio egoísta es un miembro de C. Cada persona i define una medida de intervalo $v_i$ que representa el valor que tiene para ella cada uno de los miembros de C; así, para cada principio p perteneciente a C, $v_i(p)$ es la evaluación que i hace de p. Sea v(p) el conjunto de todas las evaluaciones individuales $v_i(p)$; lo llamaremos la *representación de valor* de p. Por último, sea V(C) el conjunto de todas las representaciones de valor v(p); en otras palabras,

$$V(C) = \{v(p);\ p \text{ es un miembro de } C\}$$
$$v(p) = \{v_i(p);\ i \text{ es una persona}\}$$

La aceptabilidad de un principio de decisión cooperativa depende, en primer lugar, de que proporcione a cada persona un mayor valor esperado que el egoísmo. Una vez garantizado esto, podemos suponer que su aceptabilidad dependerá de las comparaciones que establezcan los individuos entre sus evaluaciones de los distintos principios alternativos. En otras palabras, nuestro problema es seleccionar un principio r sobre la base de sus representaciones de valor v(r), en relación con los miembros del conjunto de representaciones de valor V(C), y el requisito especial de que para todos los individuos i, $v_i(r)$ es mayor que el $v_i(e)$, lo cual podemos escribir como que v(r) sea mayor que v(e). Pero resulta que este problema es isomorfo al problema de la negociación estudiado por la Teoría de Juegos: determinar un resultado (definido en términos de sus valores) como un punto de acuerdo mutuo, dado un conjunto de resultados posibles y un resultado fijo que representa el punto de no acuerdo.[38]

38. Una explicación del problema de la negociación puede verse en Luce, R.D. y Raiffa, H., *Games and Decisions*, Nueva York, Wiley, 1957, págs. 124-126.

En la sección 2 de la primera parte, diferencié mi idea sobre la relación entre la teoría moral y la decisión racional de la de John Rawls. Dije que los principios morales son principios *para* la decisión —empleados para elegir entre acciones o estrategias posibles—, mientras que Rawls trata los principios de la justicia como objetos *de* una elección. Sostuve también que los principios morales están en relación con la racionalidad estratégica —con situaciones en las que cada persona decide con base en las expectativas sobre las decisiones de otros—, mientras que Rawls hace depender los principios de la justicia de la solución de un problema de racionalidad paramétrica en el cual se supone que las circunstancias de la decisión son inciertas, pero fijas. Ahora podemos ver que la distinción entre mi enfoque y el de Rawls es más complejo de lo que sugerí antes. Los principios morales son, ciertamente, principios *para* la decisión, y para la decisión estratégica. Son principios para la decisión en el marco de la cooperación. Pero también son objeto de elección, en el sentido de que los principios morales serían *acordados* por personas racionales, tras analizar posibles alternativas al principio egoísta para aquellas situaciones en que el conflicto entre estrategia y pago hace la cooperación deseable. Son los principios de decisión que serían elegidos por todas las personas racionales en las situaciones en que el uso general de un principio egoísta sería perjudicial para todos.

Pero aunque los principios morales sean, no sólo principios de decisión, sino también objetos de elección, siguen sin ser objeto de una elección paramétrica. He defendido que el problema de seleccionar principios morales es isomorfo al problema de la negociación, una de las cuestiones fundamentales de la teoría de la decisión estratégica. Me parece sorprendente que, dado el papel de los principios morales en la interacción, los teóricos que han intentado relacionar la teoría moral con la teoría de la decisión racional, como John Rawls y John Harsanyi, no hayan reconocido que la teoría de la negociación, del acuerdo estratégico, ofrece el punto de enlace más apropiado.[39]

39. Rawls sí defiende su uso de la decisión individual en vez de la negociación, en la secc. 24 de *A Theory of Justice*. Harsanyi ni siquiera considera que la ética y la negociación pudieran tener alguna relación.

Al proponer que consideremos los principios morales como el resultado de una negociación racional, no estoy sugiriendo que la moralidad dependa de las habilidades negociadoras, tal como son entendidas normalmente. No hay duda de que si los negociadores fueran personas reales, los principios resultantes reflejarían sus distintas capacidades, y las ventajas o desventajas iniciales que cada uno pusiera sobre la mesa de negociación. Pero, aunque los principios morales han de ser, lógicamente, aplicados por personas de carne y hueso en sus interacciones reales con los demás, la negociación mediante la cual supongo que se seleccionan, no es real. Debemos abstraernos de la situación real de los individuos de carne y hueso, en un doble sentido. En primer lugar, como los principios elegidos han de servir de criterio para evaluar las prácticas e instituciones sociales, deben ser elegidos desde una posición *anterior* a la estructura social existente. Ha de pensarse que los individuos eligen los principios para su interacción *ex ante*, de modo que no pueden negociar desde las ventajas o desventajas particulares en que les ha puesto el funcionamiento de la sociedad. Cada uno puede negociar solamente a partir de su dotación natural. En segundo lugar, la elección de principios ha de ser determinada por las reglas de la teoría de la negociación, más que por una negociación real entre agentes imperfectamente racionales, de modo que cada persona es representada en la mesa de negociación por un yo idealmente racional, y no surgen en ese contexto diferencias debidas a las distintas habilidades negociadoras. Los principios morales son aquellos que acordarían nuestros «yoes» racionales *ex ante*, para regular nuestras interacciones cooperativas.

No es necesario suponer, sin embargo, que los yoes idealmente racionales existan tras un velo de ignorancia. Aquí —como a lo largo de mi argumento— la racionalidad es instrumental. El yo racional de cada persona está completamente informado sobre sus capacidades e intereses. Así, la idea de una negociación satisface la condición —expuesta pero, según mi opinión, no observada por John Rawls— de que la teoría moral «tome en serio la distinción entre las personas».[40] Una negociación racional lo es desde el punto de vista de cada una de las personas que participan; la negociación que determina los principios morales es así, *ex ante*, racional para todas las personas. La demanda de que la teoría moral sea parte de la teoría de la decisión ra-

40. Rawls, J., *op. cit.*, pág. 27.

cional, no centra la moralidad exclusivamente en la consideración del individuo como persona moral libre e igual,[41] sino que tiene en cuenta, además, toda la riqueza de sus talentos e intereses, sus capacidades y preocupaciones, todo aquello que le hace distinto a los demás.

8. Sostengo así que nuestro argumento reivindica la moralidad apelando directamente a cada uno de nosotros. No sólo nos va mejor a cada uno de nosotros disponiéndonos a cooperar condicionalmente con los otros, sino que los términos de esa cooperación son además determinados por un acuerdo en el que participamos todos. No importa que cada uno de nosotros comience como un egoísta, tratando de que le vaya lo mejor posible y suponiendo que este objetivo debe guiar cada decisión, cada acción. El argumento que he expuesto en esta páginas debería llevarnos a rechazar el egoísmo. Sin abandonar el objetivo de que le vaya lo mejor posible, cada uno de nosotros debe reconocer que trasladar directamente este objetivo a un principio de decisión es contraproducente. La moralidad de la cooperación condicional traslada correctamente el objetivo del egoísta a la acción.

¿Qué debe hacer, pues, un egoísta? Evidentemente, debe convertirse en un cooperador, y aceptar la moralidad. Aquí sólo he iniciado la tarea de explicar esa moralidad apelando a la teoría de la decisión racional. Mi tarea principal ha sido en realidad preliminar: describir al egoísta incompleto de modo que, viendo sus carencias, podamos buscar mejor al moralista completo.

41. La frase «Persona moral libre e igual» procede de Rawls; constituye uno de los temas centrales de «Kantian Constructivism in Moral Theory», *Journal of Philosophy*, 77 (1980), págs. 512-572.

# ASEGURAR Y AMENAZAR*

## I

¿Qué relación hay entre mi propósito (*aim*) y mis razones para actuar, decidir, o elegir? Supongamos, sólo por comodidad expositiva, que mi propósito es que la vida me vaya lo mejor posible.[1] Por sí mis-

* Distintas versiones de este ensayo han sido leídas en varias universidades de Estados Unidos y Australia, y las discusiones subsiguientes han conducido a significativas modificaciones, al igual que los comentarios escritos de Annette Baier, John Broome, Richmond Campbell, Gregory Kavka, Christopher Morris, Howard Sobel y dos lectores anónimos. Las sucesivas revisiones de este ensayo fueron escritas durante un período como *Visiting Fellow* en el programa de Historia de las Ideas de la Escuela de Investigación en Ciencias Sociales de la Universidad Nacional Australiana; así como durante un período de residencia en el Estudio y Centro de Conferencias Bellagio, de la Fundación Rockefeller, como *Fellow* de la Fundación en memoria de John Simon Guggenheim; agradezco este apoyo.

1. La idea de comenzar una discusión de la racionalidad deliberativa concretando el propósito del agente proviene de Derek Parfit, *Reasons and Persons*, Oxford, Clarendon Press, 1984, pág. 3. Pero, en lo que sigue, ignoro en gran medida su discusión de las teorías contraproducentes (*self-defeating*) en el cap. 1, a pesar de que me ocupo de un enfoque de la deliberación que la presentarían como contraproducente en relación con su propósito. Exploro algunas de mis diferencias con los puntos de vista de Parfit en «Rationality and the Rational Aim», en J. Dancy (comp.), *Reading Parfit*, Oxford, Blackwell, 1997 (págs. 24-41). También debo mencionar que no intento en este ensayo discutir o relacionar mis argumentos con el estudio reciente más importante sobre intención, deliberación y racionalidad, el de Michael E. Bratman, *Intentions, Plans and Practical Reason*, Cambridge (Mass.), Harvard University Press, 1987; o con el estudio fundacional sobre los modos de la elección secuencial de Edward F. McClennen, *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge, Cambridge University Press, 1990. Sobre algunos puntos de contacto con ambos, véase mi «Commitment and Choice: An Essay on the Rationality of Plans», en *Ethics, Rationality, Economic Behavior* (comp.) Francesco Farina, Stefano Vannucci y Frank Hahn, Oxford, Oxford University Press, 1996.

mo, este propósito no me será de gran ayuda para decidir qué hacer. Tendré que detallar un poco más qué significa que mi vida vaya bien. Para ello habrá que tener en cuenta mis distintos intereses concretos, tanto aquellos constantes durante el curso de mi vida, como otros menos duraderos. Pero, por lo que ahora nos importa, esos intereses pueden ser cualesquiera. Y, de hecho, debo recalcar que nada en el argumento que voy a desarrollar requiere que mis intereses se centren en mí mismo; mi finalidad podría ser que el mundo, visto desde mi perspectiva, fuera lo mejor posible.[2]

Mi objetivo no es discutir cuáles son mis razones concretas para una acción determinada, sino reflexionar sobre el tipo o tipos de razones de que dispongo, en general, dado mi propósito (*aim*). El carácter de mis razones depende de ese propósito. Es natural suponer que la dependencia es directa, de modo que al deliberar sobre qué hacer o elegir, simplemente pienso en cómo afectarían a mi vida, para bien o para mal, las distintas acciones que puedo realizar. Dado que mis acciones no pueden afectar a cómo ha ido mi vida antes de realizarlas, tendré en cuenta su influencia en mi vida futura —desde el momento de su realización en adelante—. Formo así expectativas sobre las consecuencias para mi vida de cada una de mis acciones posibles. Tales expectativas constituyen mis razones,[3] y la acción (o acciones) apoyada(s) por las mejores razones, será(n) —de entre todas las acciones factibles— aquélla(s) que, en el momento de su realización, forme(n) parte de, o conduzca(n) a, mi mejor vida posible.

2. Aunque mis intereses no tienen por qué estar centrados en mí mismo, sí tienen que estar basados en mí mismo; por eso mi propósito no podría ser simplemente que el mundo marchara lo mejor posible. Porque se podría suponer que si todas las personas compartieran este mismo propósito, y estuvieran perfectamente informadas, estarían de acuerdo sobre qué estados de cosas serían preferibles en relación con su propósito. Y a mí me preocupa principalmente la deliberación en situaciones en que las personas están en desacuerdo —en las que el resultado que satisfaría mi propósito no es el que satisfaría el tuyo—. Quienes piensan que este tipo de desacuerdo muestra alguna imperfección o falta en, al menos, una de las partes, no encontrarán lo que sigue particularmente interesante.

3. Y en la medida en que son expectativas razonables, constituyen buenas razones. Excepto que diga o sugiera otra cosa, supondré que las expectativas son razonables. Pero las expectativas razonables pueden ser erróneas; en parte, eso es lo que hace que surjan algunos —sólo algunos— de los problemas de la racionalidad deliberativa que examinaré.

Pero, como ya he dicho en otros lugares, este tipo de explicación no funciona.[4] Frecuentemente, la acción que mejor sirve a mi propósito es aquélla fundada en consideraciones sobre las consecuencias de mi actuar —es decir, la que formaría parte de mi mejor vida posible—. Pero frecuentemente no es siempre, y una teoría sobre la determinación de mis razones en función de mi propósito aspira a ser general. No alcanzaré en este estudio una conclusión completamente generalizable, pero sí espero indicar algunas de las complicaciones de las que una teoría explicativa satisfactoria debe hacerse cargo.

## II

Comenzaré con la adaptación de una historia, original de Hume, que empleo habitualmente.[5] Mi siembra estará lista para segar la semana que viene; la tuya en quince días. A ambos nos conviene segar juntos; es mejor que hacerlo solos. Tu me ayudarás la semana que viene si esperas que, a cambio, yo te ayude dentro de quince días. Supongamos que tú, en efecto, me ayudas. Reflexionemos ahora sobre mi decisión de ayudarte. He conseguido lo que quería: tu ayuda. Si no intervienen

4. He estado defendiendo esto, implícita o explícitamente, desde mi artículo «Reason and Maximization» (*Canadian Journal of Philosophy*, 4 [1975], págs. 411-433). Mis argumentos presentes están basados en (pero difieren en ocasiones significativamente de) la discusión del capítulo 6 de *Morals by Agreement*, Oxford, Clarendon Press, 1986, en esp. págs. 182-187 (trad. cast.: *La moral por acuerdo*, Barcelona, Gedisa, 1994); también se basan en lo que he dicho en «Deterrence, Maximization and Rationality», *Ethics*, 94 (1984), págs. 474-495; «The Unity of Reason: A Subversive Reinterpretation of Kant», *Ethics*, 96 (1985), págs. 74-88; «Reason to be moral?» *Synthese* 72 (1987), págs. 5-27; «Hobbes's Social Contract», en G.A.J. Rogers y Alan Ryan (comps.), *Perspectives on Thomas Hobbes*, Oxford, Clarendon Press, 1988, págs. 125-152; «War and Nuclear Deterrance», en S. Luper-Foy (comp.), *Problems of International Justice*, Boulder, Westview, 1988, págs. 205-221; «In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)», *Proceedings of the Aristotelian Society* 89 (1989), págs. 179-194; y «Economic Man and the Rational Reasoner», en James H. Nichols y Colin Wright (comps.), *From Political Economy to Economics —And Back?*, San Francisco, ICS Press, 1990, págs. 105-131. También están relacionados con los dos artículos mencionados en la nota 1. Sólo haré referencias específicas a mis artículos anteriores para resaltar diferencias explícitas con mi posición actual.

5. Véase David Hume, *A Treatise of Human Nature*, Oxford, Clarendon Press, 1988, págs. 520-521.

otros factores directamente relevantes, ayudarte ahora significa un coste para mí. Desde luego que ayudarte ahora sería mejor para mí que haber tenido que cosechar solo por no comprometerme a ayudarte, pero mejor aún sería haber recibido tu ayuda, y ahora no devolvértela. Este cálculo puede parecer miope. ¿Qué pasaría el año que viene?, y ¿qué hay de mi reputación? Si no te ayudo, seguro que tendré que segar solo en los años venideros, y seré mal visto por nuestros vecinos. Pero da la casualidad de que pienso vender mi granja y jubilarme en cuanto coseche; en concreto, me iré a vivir a Florida, donde es difícil que me encuentre con alguien de nuestra comunidad. Podría darse la circunstancia de que me inspirases un cierto afecto, de modo que nunca pensaría en aprovecharme de ti; pero supongamos que no es el caso —supongamos que me eres indiferente—. Entonces, si mis razones para decidir qué hacer han de derivarse de mi propósito (*aim*), tendré mejores razones para no ayudarte, dado que eso es lo más beneficioso según mi propósito.

Como somos personas racionales, ambos sabemos esto. Cada uno de nosotros podría figurarse la situación que he descrito —y cada uno de nosotros sabe que es tal como la he descrito—. No tendría sentido fingir que es diferente. De modo que tú sabes que no corresponderé a tu ayuda y, como no eres ningún primo, te quedarás tan fresco viendo cómo yo cosecho solo. Ninguno ayudará al otro y la situación será, para ambos, peor de lo necesario. Nos quedaremos sin el beneficio potencial de la cooperación.

He diagnosticado anteriormente este tipo de situación —y mi diagnóstico fue que en la situación descrita yo malinterpreto mis razones—. Si mi propósito es que mi vida vaya lo mejor posible, entonces no debería derivar todas mis razones para actuar directamente de ese propósito, pensando sólo en qué acción tendrá las mejores consecuencias. Porque si siempre delibero de ese modo, no me irá lo mejor posible. Veamos por qué. Supongamos que yo comienzo la deliberación pensando en nuestra inminente cosecha y me doy cuenta de que merece la pena que me ayudes incluso al precio de tener que ayudarte a cambio, dado que la alternativa es cosechar solo. Y me doy cuenta de que tu me ayudarás si y sólo si crees que ayudándome conseguirás mi asistencia mientras que negándomela tendrías que renunciar a ella. Me interesa, por tanto, que creas que ayudarme es la forma de conseguir mi ayuda. Es muy posible que tú creas esto si te aseguro

sinceramente que te ayudaré y que no lo creas si no soy sincero, o si no te lo aseguro en absoluto.[6] Sobre todo si suponemos que tú eres buen juez de mi sinceridad y yo un mal farsante. Así, juzgo que me merece la pena asegurarte sinceramente que te devolveré tu ayuda, a pesar de que, como consecuencia, termine ayudándote cuando llegue el momento. Debo, por tanto, ofrecerte sinceramente esa seguridad si quiero que la vida me vaya lo mejor posible.[7]

Pero si mis razones para la acción se derivan directamente de mi propósito, y soy consciente de ello, y soy consciente también de que soy racional (o eso creo), no puedo darte tal seguridad. Si tomo mis razones directamente de mi propósito, no tendré una razón suficiente para corresponder a tu ayuda, y seré consciente de ello. Y no puedo asegurarte sinceramente algo que sé que tendré razón suficiente para no hacer y que, siendo como soy una persona racional, creo que no haré. No puedo darte sinceramente esa seguridad a no ser que tenga alguna manera de volverme irracional en el momento de la decisión; pero no valdrá cualquier forma de irracionalidad, sino sólo una concreta: aquella cuyo resultado sea que yo elija lo que ahora sé que tendré razón suficiente para no elegir. Enseguida retornaré a la idea de volverse irracional pero, al margen de ella, la conclusión por ahora es que si tomo todas mis razones para actuar directamente de mi propósito, entonces tengo que resignarme a cosechar solo, y mi

6. Por supuesto que en ocasiones será efectivo asegurar en vano, y en ocasiones es de esperar que me vaya mejor si soy hipócrita que si soy sincero. Pero no siempre. En este ensayo me centro en las situaciones en que un agente espera razonablemente que la sinceridad promueva su propósito mejor que el engaño. Supongo que tales situaciones ocurren con la suficiente frecuencia como para merecer atención. También supongo, aunque no puedo defenderlo aquí, que la necesidad de ser sincero no es signo que debilidad, imperfección, incapacidad o irracionalidad por parte del agente cuyo propósito es que la vida le vaya lo mejor posible. Por otro lado, la calificación moral de la sinceridad y el engaño, y la luz que esa calificación pueda arrojar, en su caso, sobre la relación entre moralidad y racionalidad, son cuestiones que sobrepasan por completo el punto que estamos discutiendo en este momento.

7. ¿Por qué hablo de asegurar? ¿Por qué no una promesa? Quiero evitar tanto la relativa especificidad del término como las connotaciones morales que la filosofía ha conferido al acto de prometer. Asegurar algo es, en el sentido en que uso el término, más que expresar una (simple) intención, pero menos que invocar las consideraciones morales incorporadas a una promesa. El presente ensayo trata de la racionalidad, no de la moralidad.

vida no irá tan bien como podría, que es justo lo contrario de lo que pretendo.

Pero ¿es esto así? Se podría objetar que me va lo mejor posible, teniendo en cuenta que soy una persona racional. A otra persona le podría ir mejor porque podría asegurar algo que yo no puedo. Pero eso no significa que yo sea incapaz de optar a lo máximo, dentro de mis posibilidades. Lo mejor que soy capaz de hacer y, por tanto, lo mejor que mi vida puede ir, depende de la clase de persona que soy. Cosechar solo es lo mejor que puedo hacer, dado que soy una persona racional que no tiene nada que ganar de ayudarte con tu cosecha. La situación descrita es de tal tipo que las personas con intereses a largo plazo, o para quienes tiene gran valor mantener su palabra, o quienes se preocupan por sus semejantes, o quienes tienen mucho talento engañándolos, obtendrán más beneficio que las personas como yo, de las que nada de lo anterior es verdad. ¿Por qué iba a ser esto sorprendente o problemático?

Sin embargo, yo pienso que debería serlo. Ser una persona racional, en el sentido de ser alguien que actúa de acuerdo a sus razones mejores o de más peso, no determina lo que uno puede hacer en el mismo sentido que las capacidades o los rasgos de carácter. Una persona racional no es alguien para quien sólo la acción apoyada por las mejores razones es posible, sino, más bien al contrario, alguien que, entre muchas acciones igualmente posibles, selecciona la que resulte apoyada por las mejores razones. Para una persona racional es posible toda acción que realizaría (o al menos podría realizar) si tuviera una razón suficiente para elegirla.[8]

Si el tipo de razones que tiene una persona se deriva de lo que esa persona pretende, entonces sí es sorprendente y problemático que, actuando de acuerdo a sus razones, deba esperar ocasionalmente realizar sus propósitos en menor medida de lo que podría. Si pretendo que la vida me vaya lo mejor posible, y actúo de acuerdo con las razones determinadas por ese propósito, ¿no debería esperar que me fuera, de hecho, lo mejor posible? Y sin embargo, si la teoría ortodoxa sobre la conexión entre propósitos y razones fuera correcta, re-

8. El paréntesis es necesario porque una persona podría no realizar la acción elegida, sin que eso signifique que era una acción imposible. Evidentemente, aquí se esconden complicaciones que (por suerte) quedan fuera del ámbito de este ensayo.

sultaría que a veces no debería esperar que actuar eficazmente de acuerdo con mis razones condujese a que mi vida fuera lo mejor posible. En consecuencia, propongo repensar esa conexión. Caracterizaré la deliberación racional de tal modo que el nexo entre la actuación conforme a razones y la realización de lo que el agente pretende sea más fuerte de lo que da a entender la simple suposición de que las razones justifican la acción capaz de producir, en el momento de su realización, el mejor resultado posible, en términos del propósito del agente. De todas formas, al final tendré que concluir que el hecho de que la deliberación se lleve a cabo desde varias perspectivas temporales, hace imposible que la relación entre razones y propósitos sea tal que siempre se pueda esperar que actuar de acuerdo a las razones garantiza al agente la mejor vida posible. Pero ahora estamos lejos de ese final; mi tarea inmediata es pensar cómo debo deliberar para ser capaz de ofrecer, en general, seguridades sinceras.

## III

En la situación que estamos considerando, lo mejor que puedo hacer es asegurarte sinceramente que, si me ayudas a cosechar, yo te ayudaré a cambio. Darte esta seguridad forma parte de mi mejor vida posible. Pero para darte esta seguridad he de tener una razón para ejecutarla, o al menos he de imaginar que la tengo. Debo suponer que, si me ayudas, yo tendré una razón suficiente para ayudarte. Esa razón no será que ayudarte forma parte de mi mejor vida posible a partir de entonces, porque no es el caso. Pero no por ello será una razón completamente desligada de mi propósito. Porque de los cursos de acción que puedo elegir, tomados globalmente, asegurarte que corresponderé a tu ayuda y cumplir mi palabra si me ayudas, forma parte de mi mejor vida posible.

Una vida aún mejor sería, sin duda, aquella en que te diese sinceramente la seguridad de ayudarte y después no cumpliese. Pero éstas son dos cosas que, aunque puedo hacer —puedo ser perfectamente sincero al asegurarte que te devolveré tu ayuda y luego no devolvértela—, no puedo elegir al mismo tiempo. No puedo decidir asegurarte sinceramente corresponder a tu ayuda y decidir a la vez que no haré honor a esa palabra cuando llegue el momento. Además, si deci-

do darte sinceramente la seguridad de ayudarte, la misma decisión implica que debo intentar cumplir y creer que lo haré.[9]

Se puede sostener, por tanto, que mi razón para ayudarte es que hacerlo forma parte del mejor curso de acción que puedo elegir —el curso de acción que hace que mi vida vaya lo mejor posible—. Supongamos que alguien, a quien identificaré como «nuestro crítico», está de acuerdo en que ésta es, efectivamente, una razón para ayudarte; y por tanto afirma que puedo darte una seguridad sincera. Pero, continúa, no es una razón suficiente para actuar de hecho conforme a la palabra dada. Porque también hay una razón para no ayudarte: que cuando llegue el momento, independientemente de lo que haya pasado, no ayudarte es la acción que contribuye a que me vaya lo mejor posible. Nuestro crítico sostiene que ésta es una razón más fuerte, puesto que actuar conforme a ella hará que mi vida vaya mejor. Su línea argumental es, sin embargo, errónea. El error consiste en suponer que puedo asegurarte sinceramente que corresponderé a tu ayuda. Porque si mi razón para no ayudarte es más fuerte, o suficiente, debo ser consciente de ello, dado que soy racional. Creeré que no te ayudaré cuando llegue el momento y, en consecuencia, no podré asegurártelo ahora sinceramente. Y así, una vez más, cosecharé solo, y la vida no me irá lo mejor posible. La causa es que no basta aceptar la razón de que ayudarte forma parte del mejor curso de acción que puedo elegir; debo estar convencido de que es una razón más fuerte que la razón de que no ayudarte produce las mejores consecuencias para mi vida.

En vista de esto, nuestro crítico cambia de terreno. La fuerza de la razón, argumenta, no tiene por qué ser invariable a lo largo del tiempo. Inicialmente, cuando deseo asegurarte que te devolveré tu ayuda, pienso en términos de los distintos cursos de acción que pue-

9. Ni mi intención ni mi creencia eliminan la posibilidad de que pueda cambiar de opinión, ni de que tenga razones para hacerlo. Un cambio razonable de opinión puede resultar de llegar al convencimiento de que la situación en la que comprometí mi palabra no era como yo la interpreté; otras veces resulta del descubrimiento de que la situación en la cual debo cumplir la palabra dada difiere significativamente de lo que había anticipado razonablemente. En ocasiones un cambio de opinión razonable puede resultar de una alteración en mis valores y prioridades. Estas posibilidades no me interesan particularmente en el presente ensayo, pero nada de lo que digo pretende descartarlas. De hecho, la teoría de la deliberación que presentaré ofrece un modo natural de evaluar el cambio de opinión, aunque dejaré ese asunto para otra ocasión.

do elegir y concedo más importancia al hecho de que corresponder a tu ayuda sería parte del mejor curso de acción, lo cual es correcto. En ese momento es mi mejor razón para actuar, y en consecuencia puedo asegurarte sinceramente que corresponderé a tu ayuda. Más tarde, cuando debo elegir si ayudarte o no, la situación es diferente y, también correctamente, concedo mayor importancia al hecho de que devolverte tu ayuda no sería la acción con mejores consecuencias. Sin embargo, este nuevo argumento de nuestro crítico es también erróneo. Sin duda que la fuerza de una razón no se mantiene necesariamente a lo largo del tiempo, pero hay un límite para su variación, y es que su fuerza en un momento dado no resulte anulada por la conciencia, en ese momento, de su fuerza en un momento diferente. La conciencia de que lo que pesará más en mi deliberación en el momento de elegir —llamémoslo tiempo t— es que no ayudarte sería lo mejor, no es obstáculo para que ahora actúe de un modo que conceda menor peso a esa consideración; pero sí lo es para que actúe ahora de un modo que me exija conceder menor peso a esa consideración en t. Y eso es precisamente lo que se necesitaría para poder asegurarte sinceramente que te ayudaré en t.

En vano propondrá nuestro crítico un esquema que acepte simultáneamente a) que el hecho de que ayudarte no forme parte de mi mejor vida posible es una razón suficiente para no ayudarte; y b) que soy racional y consciente de mi racionalidad, y que a pesar de todo incluya la suposición de que puedo asegurarte sinceramente mi ayuda a cambio de la tuya. Nuestro crítico se empeña en defender que siempre tengo más razón para realizar la acción que, en el momento de su realización, forma parte de mi mejor vida posible, pero no quiere condenarme a cosechar solo, ¿qué salida le queda? Una posibilidad sería introducir un cambio en la situación, de modo que ayudarte tuviera las mejores consecuencias para mi vida; sin embargo, debemos excluirla, pues sólo sería un truco. En la práctica siempre sería posible hacerlo, desde luego. Pero es casi seguro que resultaría más costoso que asegurarte mi ayuda, simple y sinceramente. Además, eludiría el verdadero problema: que derivar mis razones para actuar directamente de mi propósito es en ciertas situaciones contraproducente e imposibilita, de hecho, el logro de mi propósito.

Antes de considerar el siguiente movimiento de nuestro crítico, debo detenerme para advertir que la reflexión sobre la posibilidad de

asegurar revela una dificultad en mi descripción del nexo entre las acciones o elecciones posibles y las acciones o elecciones racionales. Defendí que la racionalidad no es determinante de la posibilidad, de modo que aunque ocasionalmente un agente racional no pueda realizar cierta acción por ser contraria a sus razones para actuar, ello no implica que la acción no sea factible para él. Es cierto que la racionalidad de una acción *x* no determina su propia posibilidad, pero sí puede ser determinante de la posibilidad de otra acción *y*. Porque si la realización de *y* exige que un agente tenga la intención de realizar, o crea que realizará, *x*, pero el agente está convencido de que realizar *x* sería irracional, entonces no podría realizar *y* (mientras se considere a sí mismo un agente racional). Esto es exactamente lo que está implícito en mi discusión sobre el asegurar. Dar una seguridad sincera requiere que uno tenga la intención de realizar, y crea que realizará, la acción asegurada. De modo que las creencias sobre la racionalidad propia y las razones posibles para realizar la acción asegurada afectan a la posibilidad misma de asegurar.

## IV

Tras esta digresión, volvamos a nuestro crítico, que aún tiene un movimiento posible: la propuesta de que es racional volverse irracional. Tal propuesta sostendría que aunque es irracional ayudarte, es racional convencerme a mí mismo de lo contrario.[10] Si creo que hay una razón suficiente para devolverte tu ayuda, puedo asegurarte sinceramente que te ayudaré. Y como me va mejor si te doy esa seguridad, es racional hacerme capaz de dártela, y así poder dártela de hecho. Claro que cuando llega el momento de decidir si en realidad te ayudo o no, estaría bien —reconoce el crítico— poder hacer lo que es verdaderamente racional. Sin embargo, eso sólo podría ocurrir por una feliz casualidad. No puedo planear tener una visión de lo que es racional ahora y otra más tarde —y si pudiera, posiblemente te darías cuenta, y por muy sincera que fuese mi palabra ahora (al estar basada

10. El crítico que así argumenta podría ser Derek Parfit; véase *Reasons and Persons*, págs. 9-13. Pero no me detendré a considerar en qué medida corresponden los argumentos de Parfit a los que atribuyo a nuestro crítico.

en una cuidadosa selección de mi idea presente de lo que es racional), no serviría de nada.

Así pues, nuestro crítico argumenta que conseguir que nos vaya lo mejor posible depende de la posibilidad de creer que a veces tenemos razón suficiente para realizar una acción distinta de la que en ese momento formaría parte de la mejor vida posible. Es completamente razonable, por cierto, tener esta creencia. Se trata de una creencia sobre las razones para actuar que conduce a que la vida de uno vaya lo mejor posible, por lo tanto existe una razón suficiente para mantenerla. Pero el hecho de que haya una razón para mantener esta creencia no tiene nada que ver con el hecho de que sea verdadera; y —según nuestro crítico— es totalmente falsa.

Nuestro crítico supone que la racionalidad está determinada en su carácter por el propósito del agente (y en esto estoy de acuerdo). Puesto que pretendo que mi vida vaya lo mejor posible, lo racional es que en cada ocasión en que tenga que actuar o elegir, haga o elija lo que, juzgado desde la perspectiva de esa ocasión, más contribuya a lograr ese propósito. Ésta es la tesis fundamental de nuestro crítico; precisamente la que yo quiero refutar. Él advierte que, desafortunadamente, ser racional no es lo que mejor conduce a lograr mi propósito. Ahora bien, de esto no se sigue que ser irracional lo sea, si no se especifica en qué consiste la irracionalidad. Lo que lleva más eficazmente a lograr mi objetivo es un tipo concreto de irracionalidad, que puede considerarse un modo alternativo de deliberación. El crítico supone que una persona delibera racionalmente en la medida en que considera, en cada situación, qué acción o decisión de entre las posibles formaría parte de su mejor vida posible de ese momento en adelante. Pues bien, no es que una persona racional deba renunciar a la deliberación. Lo que ocurre es que debería deliberar considerando qué acción o decisión de entre las posibles es requerida por el curso de acción que forma parte de su mejor vida posible.[11] Esa persona puede esperar razonablemente que su vida vaya mejor si delibera de este modo que si lo hace racionalmente.

11. Valga esto como explicación, a grandes rasgos, del modelo de deliberación que propongo. De hecho, lo que nuestro crítico describiría como un modo irracional de deliberación que es racional adoptar, es lo que yo defenderé en secciones posteriores de este artículo como el modo racional de la deliberación.

Luego no siempre es racional deliberar racionalmente. (O, más precisamente, siempre y cuando alguien sea capaz de llegar a deliberar de un modo que haga que su vida vaya mejor, no es racional que delibere racionalmente. En lo sucesivo daré por sobreentendida esta salvedad.) ¿Cómo hemos de entender esta tesis? Lo primero es eliminar la ambigüedad de la idea de una deliberación racional. Una deliberación puede entenderse como racional [en cuanto al resultado] o [en cuanto al modo en que se realiza. La deliberación es racional en cuanto a su resultado si produce una acción o elección racional.] Y un modo de deliberación es racional en la medida en que procede mediante la consideración de los factores realmente relevantes en relación a lo que se pretende. Estos factores son variados; puede que deliberar según buenas reglas empíricas produzca una elección racional con más efectividad que la deliberación sobre los factores reales de los que depende la racionalidad de una elección. Al agente le importa, sobre todo, el resultado, y no el modo de su deliberación; así pues, quizá lo que mejor defina como racional a un modo de deliberación es precisamente el hecho de que garantice un resultado racional. Y en este sentido, es claro que no siempre es racional deliberar racionalmente. De hecho, en algunas situaciones puede que lo racional sea no deliberar en absoluto.

Pero no es éste el sentido en que nuestro crítico afirma que a veces no es racional deliberar racionalmente. Su tesis es que no es racional deliberar de un modo que sí produce un resultado racional: la acción o elección basada en las mejores razones reales de que el agente dispone. Si creo que cuando llegue el momento de decidir si te ayudo o no, deliberaré de un modo que produzca lo que (desde ese punto de vista) es el resultado realmente racional, entonces no puedo asegurarte sinceramente que te devolveré el favor. Por eso, nuestro crítico sostiene que es racional autoconvencerme de que ayudarte en reciprocidad es lo racional; y que debo creer que mis razones para actuar justifican que te ayude; y que es racional deliberar de cualquier modo que produzca como resultado la elección de ayudarte. Nada de esto es verdad, evidentemente. Pero si me convenzo de que lo es, puedo asegurarte que te ayudaré, y así lograré que me vaya lo mejor posible, por lo menos en lo que se refiere a esta situación. Por tanto, si puedo autoengañarme, debo hacerlo.

Al reconocer que no siempre es racional deliberar de modo racional, estamos reconociendo la imperfección de gran parte de la deli-

beración cotidiana. El mejor método para que un deliberador ideal llegase a un resultado racional sería deliberar de modo racional —esto es, en términos de los factores que hacen que su decisión sea racional—. Para las personas reales, que no somos deliberadores ideales, puede ser mejor deliberar de otros modos. Pero la afirmación del crítico —que la deliberación encaminada a producir un resultado racional puede no resultar racional— no se refiere a la imperfección de la deliberación empírica. El crítico no nos está previniendo contra nuestra debilidad como deliberadores, sino contra el fracaso de quienes deliberan racional y correctamente con el objetivo de que la vida les vaya lo mejor posible. Nos previene contra la tentación de emular a los deliberadores ideales, en circunstancias en que tal emulación sería factible.

Lo que hace aparentemente atractiva la posición de nuestro crítico es, creo, un paralelismo implícito entre acción y creencia. Creer algo es creer que es verdad. La verdad impone una condición a la creencia, y esa condición es independiente (según mi concepción realista) de los procedimientos epistémicos. Como los procedimientos epistémicos son falibles en relación con la verdad, pueden calificarse como imperfectos, en terminología de Rawls.[12] La creencia racional ha de definirse en relación a estos procedimientos epistémicos imperfectos; una creencia es racional sólo si está justificada mediante procedimientos epistémicos adecuados.[13] Si se acepta esto, tiene sentido suponer que en algunas situaciones no sería racional que el agente formase creencias racionales. Porque la racionalidad de «formar una creencia», considerada como una acción posible, está relacionada con el propósito del agente de lograr el mayor bienestar posible en su vida, y quizás alcance mejor este propósito si forma una creencia que no está suficientemente justificada mediante procedimientos orientados a la verdad. Y, en este sentido, sería comprensible la cautela, no contra nuestras debilidades como sujetos de conocimiento con ca-

12. Para la discusión de procedimientos perfectos, imperfectos y puros, en el contexto de la justicia, véase John Rawls, *A Theory of Justice*, Cambridge, (Mass.), Harvard University Press, 1971, págs. 85-86.

13. No supongo que en un contexto dado un procedimiento apropiado deba producir como máximo sólo una creencia, ni que en cada contexto dado haya sólo un procedimiento epistémico adecuado. La lógica de las creencias es otro de los temas que me alegro de no tener que afrontar en este ensayo.

pacidad de formar creencias —entre las que estarían los deseos imposibles, la confianza injustificada en la autoridad, o la credulidad—, sino contra el fracaso de quienes ven que formar racional y correctamente sus creencias no trae como consecuencia la realización de sus mejores vidas posibles. Sin embargo, puede que seamos incapaces de adoptar estrategias directas para evitar este fracaso; pues como creer algo es creer en su verdad, no podemos simplemente decidir formar una creencia que no esté suficientemente justificada mediante procedimientos orientados a la verdad. Quizá sea posible hacerlo indirectamente en algunos casos —tema que ha tratado Jon Elster (siguiendo a Pascal).[14]

Se podrían establecer distintos paralelismos entre acción y creencia, pero aquí basta con el que parte de la idea de que el propósito del agente es la condición de sus acciones. Llamaré «éxito» a esta condición; una acción tiene éxito sólo si, en el momento de su realización, forma parte de la mejor vida posible del agente. El éxito es independiente de los procedimientos deliberativos del agente, y estos procedimientos son falibles (no garantizan el éxito de la acción); son, por tanto, imperfectos. La acción racional se define en relación a estos procedimientos deliberativos imperfectos; una acción es racional sólo si está basada en procedimientos de deliberación adecuados. Pero a partir de este punto no se puede seguir manteniendo un paralelismo estricto con las creencias. Quizás alguien tenga una vida mejor si forma una creencia insuficientemente justificada mediante procedimientos orientados a la verdad, y puede que esté en condiciones de reconocerlo así. Sin embargo, aunque quizá también le vaya mejor si realiza una acción insuficientemente justificada mediante procedimientos dirigidos al éxito, es imposible que esté en condiciones de reconocerlo así en el momento de realizarla, y por tanto no puede basarse en ese reconocimiento para convencerse de que sea racional renunciar a tales procedimientos. No obstante, la base para un paralelismo parcial ha quedado establecida con la aceptación, admitida por nuestro crítico, de que en algunas situaciones se puede esperar razonablemente que al agente le vaya mejor si los demás creen que re-

14. Jon Elster, *Ulises and the Sirens*, Cambridge, Cambridge University Press, 1979, págs. 47-54; trad. cast.: *Ulises y las sirenas*, México, FCE, 1989.

chaza los procedimientos dirigidos directamente al éxito; y esa creencia de los demás puede eventualmente depender de que el agente esté sinceramente dispuesto a rechazar tales procedimientos. Si esto es así, nuestro crítico aceptará también que es racional que un agente esté dispuesto a usar procedimientos deliberativos inapropiados para lograr el éxito, y a realizar las acciones recomendadas por esos procedimientos, incluso aunque sean irracionales. En resumen: existe una condición para las acciones, el éxito, que corresponde a la condición de la creencia, la verdad; existen procedimientos imperfectos orientados por esa condición; y en algunas ocasiones hay razones de hecho para rechazar esos procedimientos.

Pero ¿por qué aceptar, con nuestro crítico, el paralelismo entre la acción y la creencia? En vez de usar como condición de la racionalidad de los procedimientos deliberativos el éxito de las acciones que determinan, ¿por qué no emplear el criterio de cómo contribuye el empleo de esos procedimiento al logro del propósito del agente? ¿Por qué no determinar su racionalidad teniendo en cuenta no sólo las consecuencias de las acciones que prescriben, sino también el efecto de aquellas que hacen posibles? Los procedimientos deliberativos que hacen posible dar seguridades sinceras contribuyen a una mejor vida para el agente, porque le permiten elegir cursos de acción que de otro modo no podría elegir; sin duda, este efecto ha de tenerse en cuenta al juzgar la racionalidad de los procedimientos.

Antes dije que a un agente le interesa sobre todo el resultado de la deliberación, y no tanto su modo, pero ahora debemos reconocer que esa afirmación es ambigua. ¿Qué interesa más al agente, el efecto general de emplear ciertos procedimientos deliberativos, o los resultados particulares que sus procedimientos recomiendan? Evidentemente lo primero, porque lo que tiene una influencia más directa en su propósito de que su vida vaya lo mejor posible es el efecto de emplear ciertos procedimientos. Por tanto, los procedimientos deliberativos son racionales si su empleo produce, como efecto, que la vida del agente vaya lo mejor posible; teniendo en cuenta que ese efecto incluye no sólo las acciones que los procedimientos determinan, sino también las que hacen posibles. Ahora sí podemos convenir con nuestro crítico que una acción es racional sólo si está basada en procedimientos deliberativos adecuados. Pero como hemos roto la conexión directa entre la deliberación racional y los resultados particu-

lares, una acción puede ser racional incluso aunque al tiempo de su realización no forme parte de la mejor vida posible del agente —sin necesidad de que el agente crea lo contrario.

Si puedo esperar que siendo capaz de asegurar algo y cumplirlo —y haciéndolo— voy a tener un resultado más beneficioso que eligiendo cualquier otro curso de acción, entonces lo mejor que puedo hacer para mi vida cuando llegue el momento de decidir si ayudarte o no, es emplear procedimientos deliberativos que no se dirijan de modo inmediato al éxito de las acciones, esto es, que no seleccionen las acciones que en el momento de su realización formen parte de mi mejor vida posible. Y tales procedimientos serán racionales, al igual que las acciones por ellos determinadas, incluso aunque no pueda esperar que satisfagan el criterio del éxito.

El supuesto paralelismo que trata el éxito en la acción como si jugara el mismo papel que la verdad respecto de la creencia es erróneo. No hay razón para suponer que exista una condición de las acciones particulares que nos facilite un enfoque adecuado sobre los procedimientos deliberativos del modo en que la verdad facilita un enfoque adecuado sobre los procedimientos epistémicos. Esto no significa negar la posibilidad de determinar veritativamente la racionalidad de las acciones. Pero, desde mi punto de vista, esa verdad depende de la relación entre las acciones y la deliberación, y la verdad sobre la racionalidad de los procedimientos de deliberación se establece a su vez determinando cuáles conducen mejor a que el agente logre lo que pretende.

Lo que nuestro crítico supone es que, aunque la deliberación deriva su racionalidad del propósito del agente, lo hace de tal modo que la deliberación racional resulta ser contraproducente para ese propósito. Incluso si un agente racional fuera capaz de volverse un deliberador irracional en el sentido apropiado —de hecho está racionalmente obligado a ello— y evitara así verse envuelto en un comportamiento contraproducente, el crítico no puede negar el carácter contraproducente de la deliberación racional; todo lo que puede afirmar es que ese carácter no implica que el intento de un agente de realizar su propósito haya de frustrarse en la práctica. Yo no digo que haya incoherencia o inconsistencia formal alguna en esta posición. El crítico explica tanto el modo en que la deliberación deriva su racionalidad del propósito del agente como el modo en que resulta luego contraproducente. Yo

únicamente sostengo que se puede dar una explicación alternativa de cómo es que la deliberación deriva su racionalidad del propósito del agente, según la cual, ésta no resulta contraproducente.[15] Según esa explicación, un agente no necesita recurrir a ninguna forma de irracionalidad para evitar que en la práctica sus deliberaciones produzcan un resultado contrario al que pretende. Y es manifiestamente erróneo tratar la deliberación racional como contraproducente si hay una explicación posible que no supone ese carácter.

## V

¿Podemos con esto despachar definitivamente a nuestro crítico? He afirmado que tengo una alternativa a su explicación de la relación entre mi propósito y mis razones para actuar; ahora tengo que justificar esta afirmación. ¿Cómo se define correctamente la deliberación racional? Veamos la siguiente propuesta. Digamos que una acción *x* es intencionalmente compatible con las acciones anteriores del agente cuando cada una de esas acciones podría haber sido realizada intencionalmente si el agente hubiera tenido la intención consciente de realizar *x*.[16] Si te aseguro sinceramente que corresponderé a tu ayuda, entonces dejar de ayudarte no es intencionalmente compatible con mis acciones anteriores; no podría haberte asegurado sinceramente mi ayuda si hubiera tenido la intención de no ayudarte. Así, los procedimientos deliberativos son racionales si y sólo si seleccionan una acción que es intencionalmente compatible con las acciones previas del agente y, entre ellas, es la que forma parte del mejor curso de acción posible. Y una acción racional es, o sería, la seleccionada usando tal procedimiento deliberativo. De este modo, sólo pueden ser racionales las acciones intencionalmente compatibles con las acciones anteriores del agente.

15. O quizá debería conformarme con sostener que se puede dar una explicación alternativa que minimiza el efecto contraproducente del cambio de perspectiva temporal (véase secc. X abajo).

16. «Podría», y no «habría sido realizada». Compro un pasaje para volar a Boston con la intención de usarlo; no lo habría comprado si no tuviera esa intención. Pero no usarlo es intencionalmente compatible con haberlo comprado; podría haber comprado el pasaje con la intención de no usarlo.

Pero esta propuesta, tal como queda establecida, es claramente insatisfactoria. Porque supongamos que absurda, pero sinceramente, te prometo que te veré mañana sin falta, olvidando que necesito tiempo para preparar una entrevista de gran importancia para mi futuro. Supongamos que nuestra reunión no es particularmente significativa, de modo que sólo te causará una pequeña molestia el hecho de que yo no acuda; está claro que es racional romper esa promesa. Pero romper mi promesa es intencionalmente incompatible con mis acciones anteriores, puesto que una de ellas fue realizar sinceramente una promesa incondicional, y por tanto no acudir es irracional según mi propuesta.

Un modo de resolver el problema de las promesas absurdas y casos similares, sería suponer que la deliberación racional debería seleccionar una acción intencionalmente compatible con las acciones anteriores del agente que fueran racionales, y diera lugar a un curso de acción al menos tan bueno como cualquier otro formado por una acción intencionalmente compatible con las acciones racionales previas. Si, [irreflexivamente,] te prometo encontrarme contigo; romper mi promesa no es intencionalmente incompatible con ninguna de mis acciones racionales anteriores. Y como romper la promesa forma parte del mejor curso de acción para mi vida, es racional hacerlo.

La propuesta original no era perfecta porque consideraba irracionales actos que son perfectamente racionales. Pero esta revisión considera racionales ciertos actos irracionales. Porque supongamos que, aunque tengo buenas razones para darte seguridad sobre mi comportamiento futuro, me doy cuenta de que lo que te he asegurado de hecho me resulta menos beneficioso, o más costoso, de lo necesario. Por ejemplo, te ofrezco una suma de dinero, digamos $100, en pago por un objeto de tu propiedad, cuando debería saber que tu estarías dispuesto a deshacerte de él por la mitad de esa suma. Una vez me has entregado el objeto, se me podría ocurrir que sería mejor si no te pagara; quizá las circunstancias sean tales que tú no podrías obligarme a pagar, ni menoscabar mi reputación si no lo hago. Desde luego que no pagarte sería intencionalmente incompatible con mi oferta de $100, pero me he dado cuenta de que no actué racionalmente al ofrecerte tanto, puesto que estaba en condiciones de saber que tú habrías aceptado $50. Así que no pagarte no sería intencionalmente incompatible con ninguna de mis acciones racionales previas, y sería mejor

para mi vida que cualquiera de las otras opciones posibles. Luego la propuesta revisada consideraría racional el acto de no pagar. Pero esto es claramente contraintuitivo.[17]

Tú estás dispuesto a venderme cierto objeto, siempre y cuando yo te pague como mínimo $50. Y yo estoy dispuesto a comprarlo incluso si tengo que pagar $100 —por eso te los ofrezco—. Pero si tú piensas que, te ofrezca yo lo que te ofrezca, sólo te pagaré realmente si al tiempo señalado para el pago considero que hice la oferta más inteligente posible, sería necio que esperaras mi pago por el solo hecho de que te hice la oferta y, por tanto, sería necio que hicieras este trato conmigo. Para obtener el objeto que deseo necesito crearte la expectativa de que te pagaré, cosa que será bien difícil si sabes que yo considero racional renegar de cualquier oferta que llegue a estimar excesiva. Y por tanto, a pesar de tu disposición a vender y la mía a comprar, no realizaríamos el intercambio. Me va peor así que si hubiera estado dispuesto a cumplir mi oferta (y tú así lo hubieras creído), incluso aunque me diese cuenta de que debería haber hecho una oferta diferente (que habría satisfecho mi propósito en mayor medida).

¿Cómo podemos distinguir los compromisos que es racional cumplir y los que no? Comencemos con la idea de que es necio asegurar algo si el agente podría esperar que le fuese mejor (aunque no necesariamente lo mejor posible) sin haber dado seguridad alguna. Es necio prometer verte mañana pase lo que pase porque podría esperar que me fuera mejor si no te hubiera asegurado nada —aunque quizá me iría aún mejor si te hubiera dado una seguridad menos categórica para poder encontrarnos de otro modo—. Pero ofrecerte $100 por algo que tu habrías vendido en $50 no es del todo absurdo, incluso aunque me habría ido mejor ofreciendo menos, porque no podía esperar que me fuese mejor si no hubiera hecho ninguna oferta en absoluto, ya que el objeto vale efectivamente $100 para mí. Según esto, es racional hacer honor a un compromiso del que se espera que al tiempo de su realización no conduzca a la mejor vida posible, sólo si no es absurdo, esto es, sólo si en ese momento se espera que cumplir conduce a una vida mejor que si no se hubiera producido compromiso alguno.

17. Además de claramente inmoral, como sin duda le parecerá al lector. Pero este ensayo trata de la racionalidad, no de la moralidad. Una vez entendamos la deliberación racional podremos investigar cuál es su relación con la moralidad.

Generalizar esta sugerencia para ofrecer una descripción adecuada de la deliberación racional no es asunto fácil. La idea clave es que los procedimientos de deliberación deberían exigir en algunos casos la compatibilidad intencional de las acciones elegidas por el agente con sus acciones previas; el problema es determinar el alcance de este requisito. Como primera aproximación para solucionarlo, consideremos lo siguiente. Supongamos que un acto es intencionalmente restrictivo en potencia sólo si el agente podría llegar a enfrentarse a una elección entre actos posibles, algunos de los cuales fueran intencionalmente incompatibles con él. Y supongamos que un acto es intencionalmente restrictivo de hecho sólo si el agente tiene que afrontar una elección entre actos posibles algunos de los cuales son, en efecto, intencionalmente incompatibles con él. Si al deliberar, el agente no identifica ningún acto en su comportamiento anterior como intencionalmente restrictivo de hecho con respecto a la elección presente —de modo que ninguno de los actos que podría elegir es intencionalmente incompatible con ninguno de sus actos anteriores—, entonces debe elegir el acto que, al tiempo de la elección, forme parte de su mejor vida posible.

Sin embargo, si el agente identifica un acto intencionalmente restrictivo de hecho, entonces tiene que proceder a comparar dos expectativas. La primera expectativa está determinada por la suposición de que la elección ha de restringirse a aquellos actos intencionalmente compatibles con los actos previos; la segunda, por la suposición de que no se hubiera realizado ningún acto intencionalmente restrictivo en potencia.[18] Se construye después el mejor escenario posible para cada caso. El agente considera qué acto es el más beneficioso, de entre aquellos intencionalmente compatibles con sus actos pasados, y se forma una expectativa sobre cómo le iría en este caso. Y, por otro lado, considera qué habría pasado, y qué actos habrían sido o serían más beneficiosos, si no hubiera realizado ningún acto intencionalmente restrictivo, y se forma la correspondiente expectativa sobre cómo le habría ido en este caso. Compara entonces

18. Esto sucede cuando, el agente imagina no sólo que no hubiera realizado el acto que ha resultado ser intencionalmente restrictivo de hecho, sino que no hubiera realizado ningún otro acto que pudiera haber sido intencionalmente restrictivo.

estas dos expectativas. O bien la primera es al menos tan buena como la segunda, o bien es inferior. Consideremos cada una de estas dos posibilidades.

Si la primera expectativa, basada en una elección entre acciones intencionalmente compatibles con los actos anteriores del agente, es al menos tan buena como la segunda, basada en cualesquiera elecciones que podría haber realizado en ausencia de actos potencialmente restrictivos, entonces el agente espera que su vida vaya mejor aceptando la restricción intencional que si nunca se hubiera impuesto tal restricción. El curso de acción que incluye realizar un acto intencionalmente restrictivo y atenerse a él es parte de una vida (esperada) mejor que aquella en que se hubiera omitido cualquier restricción intencional comparable. Como es racional seguir el mejor curso de acción, el agente debería quedarse con la elección asociada a la primera expectativa. Esto es, debería elegir la acción que crea que formará parte de su mejor vida posible, entre aquellas intencionalmente compatibles con su comportamiento pasado.

Pero supongamos ahora que la primera expectativa es inferior a la segunda. En este caso, el agente espera que le vaya peor aceptando la restricción intencional de lo que le hubiera ido si no hubiera adoptado tal restricción. La restricción ya se ha producido, desde luego. Él ya no puede seleccionar la opción u opciones que habría seleccionado de no haber realizado ningún acto intencionalmente restrictivo. Pero aquel acto ha resultado ser absurdo y, por tanto, simplemente no es racional atenerse a él. El agente debería elegir la acción que juzga mejor, de entre todas las posibles.

Nótese que el procedimiento que acabo de describir no implica que el agente compare dos expectativas y elija la acción a la que asocia la expectativa superior. La segunda expectativa se basa en la hipótesis de haber actuado de modo distinto en el pasado y, por tanto, quizá no haya ninguna elección real posible que se le pueda asociar. E incluso si de alguna manera esa elección sí fuese posible, de modo que se pudiera elegir entre los mismos actos, la segunda expectativa estaría basada en una elección tomada no en circunstancias reales, sino aquellas en las que el agente se encontraría si no hubiera realizado ningún acto intencionalmente restrictivo. Si la segunda expectativa fuera superior, el agente elegiría la opción que formara parte

de su mejor vida posible, pero esa acción no sería necesariamente la que habría seleccionado si no hubiera realizado ningún acto intencionalmente restrictivo.

De hecho, puede suceder que la mejor opción posible para el agente sea intencionalmente compatible con los actos pasados, por muy desatinados que hayan sido. Si bien no es racional atenerse a un compromiso absurdo por el sólo hecho de haberlo asegurado, puede ser racional cumplirlo a pesar de todo. Volvamos al ejemplo de antes y supongamos que, sin pensar, te ofrezco $100 por un objeto del que en realidad estás deseando deshacerte. Me arrepiento casi de inmediato, pero a pesar de que podría renegar de mi oferta en este caso, ello dañaría mi reputación en un sentido que me resultaría más costoso que pagar los $100. Y así elijo hacer honor a mi oferta, no porque hacerlo sea el mejor acto intencionalmente compatible con haber hecho la oferta, sino porque, en estas circunstancias, hacerlo es simplemente la mejor acción de que dispongo. Si, por el contrario, renegar de mi oferta no dañara mi reputación, entonces ésa sería mi mejor acción posible, y [debería elegirla.] La deliberación racional abandona la condición de la compatibilidad intencional con los actos pasados cuando la restricción intencional impuesta por aquellos actos conduce a que el resultado sea peor de lo que el agente esperaría si nunca los hubiera introducido.

Un agente que aceptara esta descripción de la deliberación racional sería incapaz de realizar deliberadamente ciertos actos que pudieran ser intencionalmente restrictivos. Si denominamos *x* al acto que puede ser intencionalmente restrictivo en relación a *y*, podemos decir que un agente que realiza *x* debe reconocer la posibilidad de una elección futura en la cual *y* es una opción y es intencionalmente incompatible con *x*. Y el agente no puede realizar *x* mientras tenga la intención consciente de realizar o elegir *y* en el caso de enfrentarse realmente con esa situación. Supongamos ahora que, en vez de tener la intención de elegir *y*, el agente simplemente la incluiría como opción en sus deliberaciones. Para hacer eso, debe pensar que es posible realizarla, y tener la intención de elegirla si sus deliberaciones le condujeran a creer que es la mejor opción. El agente quizás esté seguro de que no la considerará la mejor opción; sin embargo, la presencia de *y* en sus deliberaciones le compromete con la intención condicional de elegirla si resultara

serlo. Respecto a esto mi opinión es que, del mismo modo que no puede realizar un acto intencionalmente restrictivo de *y* si tiene la intención de elegir *y*, tampoco puede realizarlo si tiene la intención de incluir *y* en sus deliberaciones, y por tanto elegirla si resultase ser la mejor opción.

No añadiré aquí nada más sobre las implicaciones de la compatibilidad intencional en los procedimientos deliberativos racionales, aunque no dudo que serán necesarios ulteriores refinamientos de la explicación recién propuesta.[19] Pero sí quiero subrayar que, aunque una descripción adecuada resulte quizá compleja, su base lógica es muy simple: mi propósito es que mi vida vaya lo mejor posible; en consecuencia, la regla general es que delibero racionalmente si elijo aquella acción que, en el momento de su realización, espero que forme parte de, o que me conduzca a, la mejor vida posible. Pero a veces la mejor vida posible es aquella en que soy capaz de comprometerme a realizar una acción cuyas consecuencias esperadas (si la realizo) serán que no me vaya tan bien como me iría si realizara una acción alternativa. Sin embargo, es racional aceptar tal compromiso, y restringir mi deliberación subsiguiente a las acciones intencionalmente compatibles con él, dado que al hacer eso actúo de un modo que espero conduzca a que mi vida vaya mejor de lo que habría esperado razonablemente que fuera si no me hubiera comprometido. Como agente racional, no seré capaz de comprometerme a realizar aquellas acciones de las que (en el instante de aceptar el compromiso) crea que el resultado de su realización será inferior al resultado de no comprometerme; pero sí seré capaz de dar una seguridad, y cumplirla, cuando me beneficie hacerlo.

19. Por ejemplo, supongamos que un inesperado efecto secundario de un acto intencionalmente restrictivo hace que mi vida vaya peor, como consecuencia de haber realizado tal acto, de lo que habría ido si no lo hubiera realizado. Pero logro el beneficio esperado de mi acto y este beneficio excede el coste de conformarme a la restricción intencional, ¿me da el coste adicional del inesperado efecto secundario una razón para no conformarme? Este problema me fue planteado por un lector anónimo, a quien agradezco la sugerencia. Creo que las estructuras intencionales más complejas (políticas) que introduzco en la sección IX me ofrecen los recursos necesarios para resolverlo, pero dejaré esa discusión para otra ocasión.

## VI

Consideremos la conocida «paradoja del tóxico», de Kavka, a la luz de esta descripción de la deliberación racional.[20] Imagínate que un rico investigador que realiza un estudio empírico sobre la decisión depositara un millón de dólares en tu cuenta bancaria esta noche a las doce en punto, si en ese momento cree que tú tienes la intención sincera de beberte, mañana a las ocho de la mañana, una dosis de cierta droga tóxica. Tú sabes que el tóxico en cuestión te pondrá terriblemente enfermo durante un día, pero no dejará secuelas. Imagina también que el investigador y generoso donante es conocido por su sagacidad a la hora de discernir las intenciones del prójimo. Tienes pocas esperanzas de burlarle; tu expectativa es que si hoy a media noche tienes la intención sincera de beberte la droga, él creerá que la tienes y depositará el millón de dólares, mientras que si no tienes la intención sincera de hacerlo, él creerá que no la tienes y no depositará el dinero. Pero lo que de hecho hagas mañana por la mañana no tendrá efecto alguno; lo único que importa es la creencia del científico donante esta noche sobre lo que tienes la intención de hacer mañana.

Si consideras que tu vida mejorará con un millón de dólares extra, incluso aunque te cueste estar enfermo un día, tienes una buena razón para formar la intención de tomarte la droga. Es mejor ganar el millón y beber el veneno que quedarte como estás ahora. Mas, ciertamente, no quieres beberlo. Preferirías simular una convincente (pero falsa) intención de beber la droga si pudieras, pero te das cuenta de que, lamentablemente, sólo una intención sincera tiene probabilidades de convencer al futuro donante. Ahora bien, si te consideras racional, puedes formar esa intención sincera sólo si supones que mañana por la mañana no será irracional beberte la droga. Pero como todo lo que ganarás mañana por beberla es un día miserable, a muchas personas, incluyendo Kavka, les parece que no sería racional hacerlo. Desde mi punto de vista se equivocan. El resultado racional de tu deliberación mañana por la mañana es la acción que forme parte de tu mejor vida posible, siempre que sea compatible con tus com-

20. Gregory S. Kavka, «The Toxin Puzzle», *Analysis* 43 (1983), págs. 33-36. Debo advertir que Bratman y yo discrepamos sobre la resolución del puzzle; véase su *Intention, Plans and Practical Reason*, *op. cit.*, págs. 101-106.

promisos anteriores —en este caso, compatible con la intención sincera que has formado hoy de beber la droga—. De modo que la acción racional es beberla.

A algunos esto les parece completamente absurdo. ¿Cómo —preguntaría un crítico— puede ser racional beber el veneno, cuando lo único que conseguirás con eso es ponerte enfermo? Sin embargo, incluso quien esto pregunta convendrá sin duda que sería racional —a pesar del coste que pudiera tener— asegurar antes de media noche que mañana a las ocho de la mañana, si quieres como si no, ingerirás el tóxico. Supongamos, por ejemplo, que justo antes de medianoche te pudieras tender en una camilla, con las manos y pies totalmente inmovilizados, un tubo insertado en tu boca y conectado con un frasco de infusión que contuviera el tóxico. Una llave conectada a un temporizador controlaría el cierre del frasco, y quedaría programada para liberar el veneno a las ocho de la mañana. Otra llave conectada a otro temporizador controlaría tus ataduras, para que se soltaran un poco después. Si el donante aceptara este mecanismo en que te atas a ti mismo como evidencia de tu intención de beber la droga (y no como una estratagema para evitar que tu intención intervenga en el asunto), y no hubiera ningún otro mecanismo menos desagradable a mano, sin duda sería racional que emplearas éste. Salvo que fueras masoquista, pasarías una noche muy desagradable, además de un día enfermo, pero ¿no vale eso un millón de dólares? Si fuera el mejor medio para ganar el millón de dólares, sería racional infligirte una noche de miseria, además de un día de enfermedad. Pero si esto es así, sería más racional evitar la penalidad nocturna y hacerte simplemente a la idea de beber la droga; éste sería un medio mejor para conseguir el millón.

Nuestro crítico estará de acuerdo, pero insistirá en que no se trata de eso. Hacerte a la idea hoy de tomar la droga mañana es perfectamente racional —la mejor opción, si es que eres capaz de ello—. Pero beberla sigue siendo irracional, lo cual complica la tarea de hacerte a la idea. Una tarea que, de puro costosa, tal vez resulte imposible. Y al admitir esto, el crítico hace evidente el fallo que ya he denunciado en la definición de racionalidad que está presuponiendo. Al tratar la racionalidad de cada acción [discretamente] haciéndola depender estrictamente de sus consecuencias, el crítico descubre que en ciertas situaciones la racionalidad, más que una ayuda, es un

obstáculo para el propósito de realizar la mejor vida posible. Según su explicación, la paradoja del tóxico enfrenta al agente racional con la tarea de burlar su propia racionalidad. La racionalidad tiene que negarse a sí misma

El crítico piensa que es una locura beber la droga. Yo, por mi parte, pienso que la única locura es no ser la clase de persona que lo haría. Y no veo ninguna base para tratar la cordura, que aquí se manifestaría en la voluntad de beber la droga, de otro modo que como racional. Un agente que capta el nexo entre las intenciones y las acciones, y que reflexiona sobre situaciones en las cuales una intención afecta a su situación independientemente de la acción misma, comprenderá que sus razones para realizar una acción pueden derivarse de sus razones para formar la intención precedente, y no al revés. Tal agente beberá la droga. Y para él, a diferencia de Kavka y aquellos que se negarían a beber, la racionalidad producirá unas acciones que son expresión coherente de su propósito.

## VII

O eso nos gustaría. Porque lo cierto es que debo hacerme cargo de un tipo de situaciones que no parecen encajar bien en la caracterización de la deliberación racional que he propuesto: se trata de las situaciones en que intervienen amenazas. Una amenaza expresa una intención condicional: si tú actúas de cierta manera, yo responderé con un acto (u omisión) en represalia. Pero no todas las expresiones de intenciones condicionales constituyen amenazas. Para amenazarte, debo antes que nada suponer que los costes que tú esperas sufrir como consecuencia de mi represalia superan los beneficios que esperas obtener si ignoras mi amenaza, de modo que, en conjunto, prefieres ceder a la amenaza antes que ignorarla y enfrentarte a la represalia. Pero no basta sólo con eso; hacerte saber que si no me ayudas a cosechar, yo no pienso ayudarte a cambio, no es amenazarte (incluso aunque suponga que prefieres que nos ayudemos mutuamente a cosechar solo). El acto en que consiste la amenaza tiene que ser de algún modo desproporcionado como respuesta a tu conducta, sea desde un punto de vista prudencial, moral o legal. Aquí me ocupo sólo del primer caso, ya que la ley y la moralidad no son objeto de esta dis-

cusión; por tanto, me interesan las amenazas en las que la represalia es un acto simplemente costoso para mí. Esto significa que no esperaré a que el cumplimiento de mi amenaza, en caso de que llegue a producirse, tenga como consecuencia que mi vida vaya lo mejor posible. Hay que suponer, en definitiva —y sólo a efectos de este argumento—, que una amenaza (si es sincera) compromete condicionalmente a quien la emite a ejecutar una represalia que causaría: a) que la vida de la parte amenazada fuese peor que si no tuviera que soportarla; y b) que su propia vida fuese peor que si no tuviera que ejecutarla.

Para ilustrar una situación de amenaza, contaré otra historia de granjeros. Nuestras granjas están en medio del campo, y un largo camino las une a la carretera pública, quedando la tuya un poco más alejada de la carretera que la mía. En el pasado tú has mantenido la parte del camino entre ambas granjas, que sólo tú usas, y ambos hemos cooperado en el mantenimiento de la parte restante. Pero ahora tú te propones abandonar esta práctica, razonando que yo necesitaré mantener de todas formas el camino que lleva desde mi granja a la carretera, y que tú podrás usarlo sin coste alguno. Como respuesta, amenazo con dejar yo también de mantener el camino si tú no cooperas. Las condiciones que caracterizan a una amenaza se satisfacen si, como podemos suponer razonablemente, el coste que implica para ti el no tener el camino en buenas condiciones supera el beneficio que esperas al no cooperar en su mantenimiento, y el beneficio que yo obtengo de tener el camino acondicionado es mayor que el coste de mantenerlo yo solo. Si comparamos los diagramas que representan la estructura de esta situación y la situación de la cosecha, veremos claramente las profundas diferencias entre las amenazas y las seguridades (Fig. 1).[21] Se amenaza para evitar ser puesto en una situación no deseada; se asegura para conseguir acceder a una posición deseable. La amenaza tiene éxito si quien amenaza no llega a tener que realizar su compromiso de optar por una acción que no conduce a su mejor

21. Ambas estructuras corresponden a la «amenaza pura» y la «promesa pura» distinguidas por Daniel Klein en su formalización del trabajo de Thomas Schelling sobre el compromiso, pionero en este campo. Véase Daniel Klein, «A Game-Theoretic Rendering of Promises and Threats», Irvine Economics Paper nº. 90-91-21 (University of California, Irvine, 1991), y Thomas C. Schelling, *The Strategy of Conflict*, Cambridge (Mass.), Harvard University Press, 1960, caps. 2, 5 y 7.

vida posible a partir de su ejecución; la seguridad tiene éxito precisamente cuando quien asegura sí llega a la situación en que tiene que realizar esa opción a la que se comprometió.

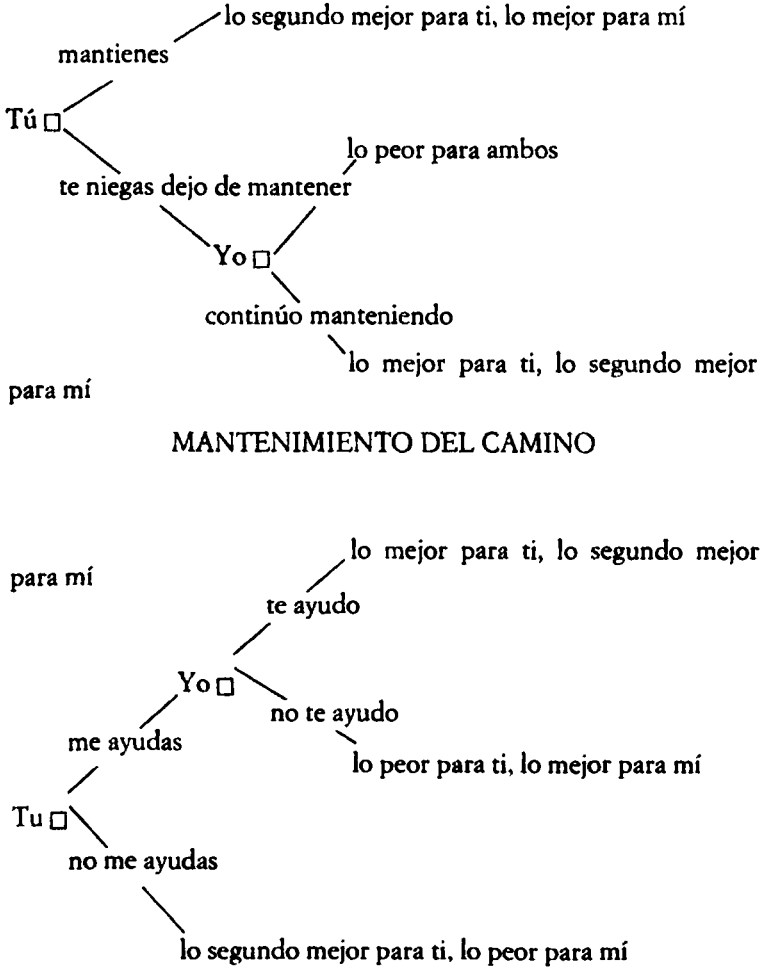



FIGURA 1

Al amenazar con no mantener nuestro camino sin tu cooperación, pretendo afectar tus expectativas sobre los costes y beneficios asociados con el hecho de dejar de compartir esa tarea, de tal modo que elijas no dejarlo. Una amenaza que logre disuadirte será lo mejor para mí y, por tanto, si puedo esperar razonablemente que mi amenaza tenga éxito, mi mejor opción parece ser plantearla. Podemos suponer que, al igual que en el caso de la cosecha no puedo esperar razonablemente que me ayudes si mi compromiso de ayudarte no es sincero, tampoco puedo aquí esperar razonablemente disuadirte de abandonar el mantenimiento cooperativo del camino si mi amenaza no es sincera. He de estar dispuesto a ejecutar mi amenaza en caso de que tú abandones, de hecho, la cooperación; esto quiere decir que tengo que pensar si sería racional que yo dejara también de mantener del camino.

Comprobemos si la deliberación racional sobre si cumplir o no una amenaza fallida es semejante a la deliberación sobre si cumplir una seguridad exitosa. Tras haber producido una amenaza sincera que no logró disuadirte, reconozco que era un acto intencionalmente restrictivo de hecho, y por tanto considero las acciones intencionalmente compatibles con él que puedo realizar. Si mis únicas opciones son seguir manteniendo el camino o dejar de hacerlo, sólo la última es compatible con mi amenaza, así que será la mejor, por defecto. A continuación comparo cómo espero que vaya mi vida si dejo de mantener el camino con cómo habría esperado que fuera si no te hubiera amenazado —o lo hubiera hecho «de farol», sin que implicara un compromiso—. Como en este caso (si no te hubiera amenazado, o lo hubiera hecho de farol) debería haber esperado que tú abandonases la cooperación, la comparación se reduce a dejar el camino sin mantener o mantenerlo sin tu ayuda. Prefiero lo segundo —de hecho, debo preferirlo si lo que planteé fue una amenaza genuina—. Así pues, creo que actuar de un modo intencionalmente compatible con mi amenaza fallida hace que me vaya peor de lo que razonablemente creo que me hubiera ido si no hubiera realizado ningún acto intencionalmente restrictivo. Obtendría un resultado mejor si no te hubiera amenazado; lo mismo que sucederá si no cumplo mi amenaza. Pero si esto es así, ¿puede ser racional cumplirla, es decir, hacer que mi vida vaya, no sólo peor de lo necesario, dadas las alternativas posibles ahora, sino peor que lo que habría ido si no me hubiera comprometido inicial-

mente? Según la descripción de la deliberación racional que propuse en la sección V, la respuesta es no. No puedo racionalmente cumplir mi amenaza y, una vez que entiendo esto, seré incapaz de pretender disuadirte de retirarme tu ayuda en el mantenimiento de nuestro camino mediante la amenaza de dejar yo mismo de mantenerlo.

Podemos generalizar a partir de este ejemplo. Al discutir los compromisos distinguí los que son racionalmente respetados de los que son necios o absurdos. Consideraba que un compromiso es absurdo cuando el agente espera que atenerse a él es más costoso que no haberse comprometido. Y proponía una teoría de la deliberación que hiciera racional atenerse a los compromisos si su concurso beneficiaba al agente sobre la hipótesis de no haberse comprometido, pero irracional hacerlo en el caso contrario, en que el compromiso habría de ser considerado absurdo. Supongamos que queremos distinguir ahora, de modo semejante (siguiendo la caracterización de la deliberación racional que he generalizado a partir del caso de los compromisos), entre amenazas que sería racional ejecutar y amenazas absurdas. Se amenaza con hacer algo desventajoso —algo que se cree producirá un efecto peor que simplemente sufrir la acción ajena que se quiere evitar mediante la amenaza—. Así, un agente debe esperar que ejecutar una amenaza, una vez abortada su pretensión disuasoria, impondrá sobre él un coste mayor que el esperado en caso de no haber planteado amenaza alguna. (De hecho, ejecutarla sólo resultará ser menos costoso si el agente se da cuenta de que había subestimado el coste de ceder, o sobreestimado el coste de ejecutar la amenaza; pero con eso no se puede contar de antemano.) Una amenaza es un acto intencionalmente restrictivo en potencia. Si es ignorada, resulta intencionalmente restrictiva de hecho; de tal modo que mantener la compatibilidad intencional con ella —cumplirla— deja al agente peor de lo que habría esperado estar si no la hubiera planteado (y no hubiera, por tanto, realizado ningún acto intencionalmente restrictivo en potencia). Y como cumplir la amenaza también deja al agente peor que no hacerlo, entonces (de acuerdo con mi teoría de la deliberación racional) no es racional cumplirla. Si aplicamos mi visión de los compromisos absurdos a las amenazas, descubrimos que las amenazas son todas ellas absurdas. Cumplir cualquier amenaza sería tan irracional como atenerse a los compromisos absurdos.

Si nuestro propósito es que la vida nos vaya lo mejor posible, nos interesaría ofrecer una seguridad, o plantear una amenaza, cuando esperásemos que hacerlo tuviera como consecuencia eso precisamente: que la vida nos fuera lo mejor posible. Sin embargo, aunque atenerse a una seguridad dada puede ser parte del curso de acción que mejor promueve ese propósito, nunca puede esperarse lo mismo del cumplimiento de una amenaza. Como razón para atenerse a una seguridad dada, se puede aducir que así es mejor que si no se hubiera dado ninguna seguridad; pero no se puede dar una razón similar para cumplir una amenaza. Sin tal razón, sería irracional hacer lo que conduce a que la vida nos vaya peor de lo necesario, y por lo tanto sería irracional cumplir la amenaza. Por otro lado, si no se puede esperar tener esa razón, no se podría realizar racionalmente el tipo de acto que es sabido que restringe intencionalmente el ámbito de elecciones a opciones que serían irracionales sin él. Como agente racional, soy capaz de ofrecer seguridades sinceras, pero incapaz, en principio, de plantear amenazas sinceras.

## VIII

Éste podría parecernos un resultado feliz: asegurar es frecuentemente racional; amenazar nunca lo es. Estaría bien que las cosas fueran así. Se recordará que comencé mi discusión de las amenazas sugiriendo que no encajaban muy bien en mi explicación de la deliberación racional, pero a lo mejor esto es una virtud de mi teoría. Personalmente, sería un poco embarazoso ya que he defendido en más de una ocasión la racionalidad de la disuasión, que es una forma de amenaza.[22] Si la teoría de la deliberación racional que acabo de exponer es correcta, entonces mi defensa de la disuasión era errónea. Puede que sea embarazoso descubrir esto, pero en modo alguno es motivo de tristeza. Lo malo es que si mi teoría de la deliberación racional fuera correcta, parece que se derivarían otras consecuencias menos agradables. He aquí un ejemplo que las muestra (véase Fig. 2).

22. Véanse «Deterrence, Maximization and Rationality», y «War and Nuclear Deterrence», ambos citados en la nota 4.

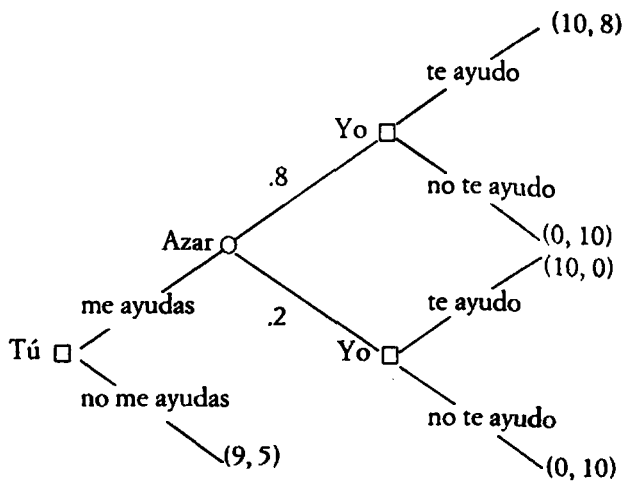



FIGURA. 2. Los números entre paréntesis representan la medida de cómo le va a cada uno; la correspondiente a ti aparece en primer lugar.

En esta situación el azar determina si la ayuda mutua me beneficia más o menos que no ayudarnos. Supongo en el ejemplo que el azar determina que la ayuda mutua es beneficiosa para mí con una probabilidad del 0,8; las probabilidades de cada uno de los resultados alternativos se muestran en el diagrama. Mi expectativa inicial es que la ayuda mutua me beneficiará; esto lo determino calculando el resultado esperado de la ayuda mutua (una suma ponderada) sobre cómo es de esperar que me vaya, que es (0,8 * 8)+(0,2 * 0), ó 6,4; y comparándolo con el resultado si tu no me ayudas, 5. Visto lo cual, parece que asegurarte que te ayudaré es racional; espero que me vaya mejor comprometiéndome y cumpliendo que no comprometiéndome.[23] Pero supongamos que me comprometo a ayudarte y tengo mala suerte; en ese caso acabaré con 0 si te ayudo. Entonces sin duda que

23. Aquí estoy empleando la teoría ortodoxa de la decisión racional en condiciones de riesgo, que implica la maximización de la utilidad esperada. Ya no estoy tan convencido como antes de que éste sea el único procedimiento racional de decisión, pero el problema que estoy examinando surge también al emplear otros procedimientos plausibles en condiciones de riesgo.

me habría ido mejor sin comprometerme, y si no es racional cumplir un compromiso que resulta dejarme peor que si no lo hubiera realizado, debo concluir que no es racional atenerme a lo asegurado. Será racional atenerme a ello sólo si ayudarte me proporciona un resultado de 8. Según esto, sólo podría ofrecerte sinceramente una seguridad condicional; te ayudaré, pero sólo si el azar me favorece, de modo que la ayuda mutua me sea más beneficiosa que si no nos hubiéramos ayudado. Pero tal seguridad condicional no será suficiente para que te decidas a ayudarme. Porque si calculas el resultado esperado de ayudarme, dada mi seguridad condicional, verás que es (0,8 * 10)+(0,2 * 0), u 8, que es menos que el resultado si no me ayudas, 9. Sólo un compromiso incondicional te impulsaría a ayudarme, al darte un resultado esperado de 10.

Sin embargo, no hay duda de que me beneficia ser capaz de darte una seguridad incondicional. Es evidente que debo revisar mi teoría de la deliberación racional para explicar este supuesto. Y la revisión parece fácil. Lo que se necesita es relacionar la deliberación posterior sobre si cumplir un compromiso o ejecutar una amenaza con la deliberación previa sobre si ofrecerlos. En mi ejemplo, calculando sobre la base de probabilidades estimadas, determino que el beneficio neto esperado de conseguir tu ayuda asegurándote la mía a cambio, es mayor que el resultado de no ayudarnos mutuamente. Por tanto, considero racional asegurarte sinceramente que te ayudaré. Cuando llega el momento de considerar si cumplir o no, lo que debo hacer es simplemente remitirme a mi deliberación anterior; si reafirmo mi juicio de que en aquellas circunstancias, tal como las conocía, era racional darte una seguridad, entonces concluyo que debo atenerme a ella, incluso aunque en estas circunstancias, tal como ahora las conozco, me habría ido mejor si nunca te la hubiera dado. De igual modo, al considerar mi amenaza de abandonar el mantenimiento del camino si tú dejas de colaborar en la tarea, estimo la probabilidad de que mi amenaza sea efectiva y calculo si, sumando el beneficio esperado de garantizar que continúes colaborando (si mi amenaza es efectiva) y el coste esperado de que el camino quede sin mantenimiento (si mi amenaza fracasa), puedo en conjunto esperar un resultado mejor que si tuviera que mantener el camino solo. Y si espero un resultado mejor en ese caso, entonces es racional amenazarte, y eso es lo que debo hacer. Si la amenaza fracasa, me remito a mi deliberación anterior; si

reafirmo mi juicio de que en aquellas circunstancias, tal como las conocía, fue racional plantear la amenaza, entonces decido ejecutarla, aunque en estas circunstancias, tal como ahora resultan ser, me habría ido realmente mejor si no te hubiera amenazado.

En el pasado he defendido esta teoría de la deliberación. He sostenido que permite al agente establecer en su vida y sus acciones una coherencia]que la teoría ortodoxa, con su requisito de que[en cada ocasión se elija la acción que en ese momento tenga el mayor valor esperado, impide. El agente que acepta el canon ortodoxo de la razón no puede, salvo indirectamente, interrumpir el carácter inexorablemente prospectivo de su deliberación para aceptar compromisos que serían claramente beneficiosos para él. La alternativa que acabo de describir permite tales compromisos, al autorizar una deliberación retrospectiva]—en la medida en que apela a intenciones o planes adoptados con anterioridad, a compromisos adquiridos y amenazas planteadas previamente—. No se trata de una deliberación automáticamente retrospectiva; el agente no ha de respetar necesariamente sus intenciones anteriores si se encuentra en circunstancias muy diferentes de aquellas que imaginó al formarlas, o si se convence de que *en aquel momento* no tenía buenas razones para formarlas. Pero el concepto subrayado nos obliga a detenernos.

Lanzo una amenaza; pero no logra disuadirte. Ahora estoy, se supone, exactamente en la situación en que imaginé que estaría si mi amenaza fracasaba. Puede incluso que haya considerado ese fracaso como probable. Porque lo que determina que la opción de amenazar tenga más valor esperado que cualquier otra alternativa no es simplemente la probabilidad de éxito, sino la suma ponderada del beneficio que representaría su éxito y el coste que representaría su fracaso. Si el éxito de cierta amenaza me proporcionaría un gran beneficio, mientras que su fracaso supondría sólo un pequeño coste, debería considerar que merece la pena plantearla, aun anticipando su fracaso. Supongamos por tanto que, habiendo fracasado mi amenaza, estoy en la situación en que esperaba estar, y resulta ser como la había previsto. Sin embargo, mis creencias se habrán alterado significativamente.[24]

24. Comienzo aquí el argumento que, según me parece, niega la racionalidad de la disuasión y, en general, del comportamiento amenazador, que sostuve en mis artículos anteriores sobre el tema, citados en la nota 4.

Porque al plantear la amenaza, creía que ése era mi mejor curso de acción posible, y ahora sé que aquella creencia era equivocada. Te amenacé basándome en mi estimación previa de la probabilidad de éxito, y quizás esa estimación era muy razonable; pero el caso es que amenazarte no era beneficioso de hecho. Yo esto no lo sabía, y por eso se puede decir que actué racionalmente al amenazarte. Yo no sabía que, de hecho, mi vida iría peor como consecuencia de la amenaza, de lo que habría ido sin ella. Si, a pesar de todo, ejecuto mi amenaza, estaría actuando, aparentemente, sin tener en cuenta este nuevo conocimiento de que mi curso de acción no tiene como consecuencia que mi vida vaya lo mejor posible. ¿Cuál puede ser, entonces, la relación entre la ejecución de la amenaza y mi propósito?

Si suponemos que un agente racional actualizará su evaluación de la racionalidad del comportamiento amenazador de acuerdo con su conocimiento presente del efecto de su amenaza sobre su propósito de que la vida le vaya lo mejor posible, concluiremos que normalmente no es racional ejecutar amenazas. Igualmente, concluiremos que no es racional atenerse a una seguridad si se descubre que hacerlo conduce a un resultado peor que si no se hubiera dado seguridad alguna. Pero no concluiremos que no es racional cumplir una seguridad en aquellos casos, más típicos, en los que hacerlo deja al agente mejor de lo que habría esperado estar si no se hubiera comprometido de ningún modo. La razón es que encontrarse en la situación en que hay que cumplir un compromiso no aporta ningún conocimiento nuevo que indique que, en contra de las expectativas, el curso de acción elegido no era el que conducía a la mejor vida posible. Si te aseguro sinceramente que corresponderé a tu ayuda, y tú me ayudas, tengo todas las razones para pensar que mi compromiso ha producido exactamente el resultado deseado: que me hayas ayudado, cosa que no habrías hecho si no te hubiera asegurado mi reciprocidad.

Así, este argumento contra la simple subordinación de la deliberación posterior a la deliberación previa, no nos devuelve a la visión estrictamente prospectiva de la deliberación que he estado tratando de refutar. Al contrario, nos devuelve a la idea de que la deliberación racional supone la valoración de la contribución de un curso de acción al propósito de que la vida vaya lo mejor posible, entendiendo que los cursos de acción se distinguen y definen por su estructura intencional. Un agente actúa racionalmente si, de entre las acciones posibles inten-

cionalmente compatibles con su comportamiento previo, realiza aquella que conduce a que su vida vaya lo mejor posible, siempre y cuando espere que le vaya mejor de lo que le habría ido si no hubiera realizado actos intencionalmente restrictivos relevantes para esta elección. Y así, parece que no sería racional cumplir amenazas, ni compromisos que, por muy grande que fuera el beneficio esperado al efectuarlos, trajeran finalmente como consecuencia que el agente acabara peor que si no los hubiera realizado. Y si, de hecho, esto fuera irracional, un agente racional no podría consciente y sinceramente efectuar esas amenazas ni ofrecer esos compromisos; sería incapaz de realizar los actos intencionalmente restrictivos en potencia necesarios para ello.

## IX

Al realizar un acto intencionalmente restrictivo en potencia, el agente crea una estructura intencional para su conducta futura. Hasta este momento he considerado dos formas muy simples de estructura intencional: seguridades y amenazas. Tomando como ejemplo el propósito de que la vida vaya lo mejor posible, he mostrado que las estructuras intencionales crean problemas para la teoría ortodoxa de la deliberación, que sostiene que las acciones racionales son aquellas que promueven el propósito del agente. He propuesto una teoría alternativa de la deliberación, que evita algunos de estos problemas aunque, como acabo de defender, parece deducirse de ella que el agente es incapaz de amenazar sinceramente, y que ofrecer seguridades sinceras es racional sólo cuando el compromiso adquirido no le obligue a arriesgarse a tener que elegir una opción cuyo resultado sea peor que la situación en que estaría si no hubiera ofrecido ninguna seguridad. Esto parece limitar seriamente lo que considero racionalmente posible; sin embargo esta sensación se debe sólo a que me he circunscrito a un espectro muy concreto de estructuras intencionales.

Además de las seguridades y amenazas particulares, un agente puede tener políticas que le exijan dar seguridades o plantear amenazas si se encuentra en algunas circunstancias prefijadas. Estas políticas son intencionalmente restrictivas en potencia. Una compañía puede adoptar sinceramente la política de tomar represalias contra sus competidores que inicien una guerra de precios sólo si tiene la in-

tención de llevar a cabo las acciones pertinentes. Dejar de represaliar la bajada de precios de un competidor sería intencionalmente incompatible con esta política del mismo modo en que dejar de ayudarte con la cosecha sería intencionalmente incompatible con la seguridad que te di. Puede ser que en un caso particular la represalia sea desventajosa para la compañía, incluso teniendo en cuenta consideraciones a largo plazo, como los efectos en su reputación. Sin embargo, la compañía, o sus ejecutivos, pueden creer razonablemente que la política de represalias es beneficiosa en conjunto, de modo que la represalia en este caso deja a la compañía mejor de lo que habría esperado estar si nunca hubiera adoptado una política que le impone realizar actos intencionalmente restrictivos. Así, dada la teoría de la deliberación que he propuesto, sería racional que la compañía tomara represalias.

Pensemos en un agente que adopta la política de amenazar y cumplir sus amenazas. Planea lanzar las amenazas que maximicen su expectativa de beneficio, teniendo en cuenta el efecto conjunto del beneficio esperado de la disuasión y el coste esperado de ejecutar las represalias. Así establece una estructura intencional para su futuro comportamiento amenazador. Y forma la intención general de ejecutar todas sus amenazas, creando así un contexto intencional en el que se enmarca cada una de sus amenazas, que se suma a la intención que el hecho concreto de su planteamiento pueda implicar. Enfrentado al fracaso de una de sus amenazas disuasorias, considera si le va mejor ejecutándola o no habiendo creado nunca la estructura intencional de que forma parte. Debe considerar no sólo cuál sería su expectativa si no hubiera planteado esta amenaza concreta, o lo hubiera hecho «de farol». Porque su política no sólo exige el cumplimiento de amenazas, sino su mismo planteamiento. Si su amenaza forma parte de una política general, no cumplirla es intencionalmente incompatible con esa política. Así, aunque ejecutar la amenaza fallida no logre que su vida vaya mejor que si no hubiera amenazado, no hará necesariamente que su vida vaya peor que si no hubiera adoptado esa política general. Los beneficios logrados, o esperados, de su política amenazadora, pueden muy bien dar lugar a que su vida vaya mejor que si no hubiera adoptado tal política, incluso aunque deba afrontar los costes de ejecutar sus amenazas fallidas.

Si las amenazas o seguridades concretas de un agente se enmarcan

en una política general de amenazar o asegurar, ese agente puede exponerse racionalmente a un riesgo mayor que si forma intenciones sobre su comportamiento futuro caso por caso. Puede amenazar sinceramente, o asegurar tomando mayor riesgo, porque sabe que será racional cumplir su amenaza o seguridad en tanto tenga, en el momento de ejecutarla, la expectativa de un resultado mejor que si no hubiera adoptado ninguna política potencialmente restrictiva de sus intenciones. Pero para que su política supere la prueba de la racionalidad debe incorporar una significativa limitación. Supongamos que, teniendo todo en cuenta, considera ventajoso plantear lo que llamaré una «amenaza apocalíptica» —una amenaza tal que, si fracasara, le obligaría a ocasionar un absoluto desastre incluso para sí mismo—. El agente piensa que la probabilidad de que la amenaza fracase es suficientemente pequeña como para que, a pesar del enorme coste que ese caso supondría, su expectativa de beneficio se maximiza amenazando. Así, su política amenazadora podría incluir el planteamiento de una amenaza apocalíptica. Mas si esta amenaza fallara, el agente se enfrentaría a un coste que superaría todos los beneficios que hubiera ganado, o que esperase ganar, mediante su política general. En ese punto, el cumplimiento de la amenaza produce como expectativa una vida mucho peor que si nunca hubiera emprendido una política de amenazar y cumplir amenazas. Y por tanto no consideraría racional cumplir su amenaza apocalíptica. En consecuencia, ninguna política racional y sincera incluiría la posibilidad del planteamiento de tal amenaza. Por muy ventajosa que una amenaza parezca de antemano, la posibilidad de plantearla sinceramente, mediante la adopción de una política que la requiera o la permita, dependerá de que el coste de su cumplimiento, en caso de ser necesario, sea superado por los beneficios totales de la política, de tal modo que adoptar dicha política sea globalmente más ventajoso que no adoptar política alguna. Un agente racional no puede sincera y conscientemente plantear una amenaza apocalíptica. La disuasión racional tiene un límite que no supe reconocer anteriormente.

Al ampliar la teoría de la deliberación expuesta en este artículo a las políticas, he podido mostrar que no establece una distinción tan radical entre seguridades (permitidas) y amenazas (prohibidas). Hay contextos en los que un agente encontraría racional adoptar una política de amenazas. Quizá se trate de contextos en que se juega con la

irracionalidad de otras personas; porque bien puede ser que la respuesta racional al eventual intimidador sea siempre una política de resistencia. En un mundo de personas totalmente racionales, quizá las amenazas resultaran completamente irracionales; quizás esto las distingue de las seguridades. Pero no puedo profundizar en estas cuestiones aquí. Me interesa sólo mostrar cómo una estructura intencional ampliada, incrementa los tipos de acciones que, aunque no tengan como consecuencia que la vida del agente vaya lo mejor posible, satisfacen los criterios de la racionalidad deliberativa.

## X

¿He logrado ofrecer una teoría de la deliberación en la que mis razones para actuar se derivan de mi propósito y me permite, si acierto a actuar conforme a mis razones, realizarlo en la mayor medida posible? ¿He logrado superar el problema que detecté en el tratamiento que da Kavka a la paradoja del tóxico (que le lleva a concluir que el agente racional tiene que burlar su propia racionalidad)? No lo parece. Porque acabo de afirmar que no sería racional para una persona ejecutar una amenaza apocalíptica, aunque esperara razonablemente que le iría mejor planteando tal amenaza. El agente racional no puede tener la intención de hacer lo que sabe que no tendrá razones para hacer, y a causa de ello es incapaz de establecer ciertas estructuras intencionales, incluso aunque esperara beneficiarse de ello. Los límites en la racionalidad de llevar a cabo intenciones implican limitaciones en la capacidad racional de formarlas. Y así, por mucho que actúe de acuerdo a mis razones, los límites que la racionalidad impone en mi capacidad intencional pueden impedirme realizar mi propósito en toda la extensión posible. Podría darse la situación en que sólo una amenaza apocalíptica me salvara de una seria desgracia y, sin embargo, me vería incapaz de plantearla. Podría darse la situación, menos dramática, en que una seguridad incondicional me beneficiara enormemente, y sin embargo fuera incapaz de darla al reconocer que, en el momento de cumplir, las condiciones podrían ser tales que el cumplimiento tendría consecuencias peores que la situación en que no hubiera establecido la estructura intencional de la que forma parte.

En vista de esto, se podría suponer que debo retornar a mi revisión anterior, según la cual la deliberación posterior se subordina sin más a la deliberación previa. Expuse mi argumento contra esta concepción en la Sección VIII; es, no obstante, una perspectiva que parece ofrecer una explicación plausible de las razones para actuar (acorde con la realización del propósito del agente en la mayor medida posible). Pero no es así. Supongamos que la deliberación posterior está subordinada a la deliberación previa, de modo que un agente puede lanzar una amenaza apocalíptica en un momento en que ésta le ofrece mejores expectativas que cualquier otra acción. Un agente que supone que tienen una razón, derivada de su propósito, para crear una estructura intencional de este tipo porque espera que la misma le conduzca al mejor resultado posible, debe también suponer que tiene una razón para actuar en coherencia con esa estructura, aun cuando reconozca que su expectativa era errónea. Pero un agente que cree tener una razón para ejecutar una amenaza apocalíptica no logra en absoluto, guiándose por esa razón, realizar su propósito. Al subordinar enteramente su deliberación posterior a la deliberación previa, persiste en un curso de acción que ahora reconoce que le lleva al desastre y, por tanto, es directamente contrario a lo que pretendía. Para poder actuar del mejor modo posible visto desde la perspectiva previa, ha de estar dispuesto a actuar en contradicción con su mejor vida posible, tal como la ve desde el conocimiento adquirido luego.

Las expectativas (razonables) sobre los beneficios y costes de diferentes acciones y cursos de acción pueden variar con el tiempo. Cualquier teoría de la deliberación racional debe tener esto en cuenta. No hay modo de describir las razones para la acción de manera que un agente pueda estar seguro de que, siguiendo fielmente sus razones, la vida le va a ir lo mejor posible) juzgada desde cualquier perspectiva temporal. He supuesto que las razones de un agente deben apoyar las acciones que realizan su propósito, considerado en el momento de la realización de la acción. Pero también he supuesto que las razones de un agente deben enmarcar sus acciones en estructuras intencionales que le permiten realizar mejor su propósito. Tratar de reconciliar ambos supuestos me ha conducido a ofrecer la presente concepción, como punto de partida hacia una teoría adecuada de la deliberación racional.

Actúo para influir en cómo irá mi vida. Cuando decido qué hacer, miro hacia adelante, a los efectos esperados de las diferentes acciones

posibles. Mis razones para actuar se derivan de esos efectos esperados. Pero cuando miro hacia el futuro, veo algo más que los efectos discretos de las acciones particulares. Que mi vida vaya mejor o peor dependerá, en parte, de las estructuras intencionales que cree. Para que mi vida vaya lo mejor posible, necesito un control sobre mis elecciones futuras que me permita dar seguridades y —al menos cuando trato con personas imperfectamente racionales— plantear amenazas. Debo ser capaz de derivar razones para la acción de las intenciones incorporadas en estos compromisos, de modo que estas razones se sumen a las razones directamente derivadas de los efectos esperados de mis acciones. Con esa idea he investigado aquí cómo pueden las estructuras intencionales capacitarme para ser «mi propio fiador» —la cita es una adaptación de Nietzsche—.[25] Una teoría adecuada de la deliberación racional sería aquella que articulase completamente lo que yo sólo he comenzado a esbozar —las relaciones entre la intención y la razón que hacen posible la realización de este ideal.

25. Véase Friedrich Nietzsche, *On the Genealogy of Morals* (1887), segundo ensayo, secc. 2, Nueva York, Vintage Books, 1967, pág. 59.

# CONTRACTUALISMO POLÍTICO

El objeto de esta investigación es la relación entre las pretensiones normativas de una sociedad y las actitudes normativas de sus miembros. Adoptaré una perspectiva contractualista, la única admisible para personas que ni esperan ni exigen que sus conciudadanos compartan su misma orientación hacia los valores y las normas. Aunque no será mi interés principal, espero contribuir también a una interpretación de las cláusulas sobre el establecimiento y ejercicio de la religión en la Primera Enmienda de la Constitución de los Estados Unidos, y sugerir algunas directrices para determinar los derechos y obligaciones derivados de la paternidad y/o maternidad. En general, intentaré responder a ciertas cuestiones sobre cómo se pueden justificar en una sociedad democrática las prácticas e instituciones sociales.

## 1

Para comenzar, describiré la perspectiva contractualista. El contractualista propone un modo de justificación de las normas que rigen la interacción. Hay que pensar que esas normas tienen un doble aspecto: por un lado amplían las oportunidades de acción, pero por otro lado limitan los modos y tipos de acción permitidos. Definen prácticas que amplían las perspectivas de los individuos, no sólo instrumentalmente, al facilitar la realización de fines establecidos con independencia de ellas, sino también intrínsecamente, al permitir que surjan fines totalmente nuevos. Pero también imponen límites, no sólo instrumentalmente, al restringir los medios que los individuos pueden emplear para alcanzar sus fines, sino también

intrínsecamente, al excluir determinadas prácticas y, en consecuencia, los fines que se pueden lograr mediante las mismas. El contractualista toma como punto de partida una situación inicial donde no hay ninguna de estas prácticas, ninguna de las posibilidades o limitaciones sociales. Partiendo de esa situación inicial, las prácticas y normas han de ser justificadas ante agentes individuales, definidos únicamente por su capacidad de deliberación racional. En este importante sentido, el contractualismo es expresión del individualismo normativo.

Aplicar el test contractualista a una norma o sistema de normas consiste simplemente en examinar si, en la hipótesis de que las personas afectadas por la norma o sistema normativo, estuvieran en situación de determinar las condiciones de su interacción, tendría sentido para cada una de ellas (desde el punto de vista de una deliberación racional) acordar unas condiciones de las que esa norma o sistema normativo pudiera considerarse expresión. Así, el contractualista sostiene que las instituciones y prácticas sociales —y las oportunidades y límites, derechos y obligaciones que crean— se pueden justificar en términos de un acuerdo hipotético; y sólo en esos términos. No es que el contractualista suponga que siempre es posible apelar a un acuerdo hipotético para satisfacer las necesidades de justificación; pero sí mantiene que tal recurso ofrece la mejor defensa posible para un sistema de regulación normativa de la interacción que sea beneficioso para todos. Si bien esto no le impide reconocer que todos los sistemas reales son, con mucho, anteriores a las personas reales que disfrutan sus beneficios y han de cumplir sus normas.

La posición contractualista es claramente minimalista desde el punto de vista normativo. El contractualista sostiene que el único espacio no controvertido para la determinación normativa de acciones es el que proporciona la deliberación individual. Lo que caracteriza a los seres humanos como animales deliberativos es: a) la capacidad de ser motivados por representaciones de estados de cosas tal como son y tal como podrían llegar a ser a través de las acciones posibles del agente; y b) la capacidad de, primero, preguntarse si tiene sentido ser motivado a la realización de un acto por cierto conjunto de representaciones; y, segundo, ser motivado de hecho por la respuesta a dicha cuestión. La idea de lo que tiene sentido hacer es el elemen-

to normativo esencial en la deliberación, y el contractualista se basa en ella.[1]

Además de minimalista, el contractualismo es constructivista en un sentido importante. El contractualista piensa que toda la estructura normativa de una sociedad depende de la normatividad deliberativa de sus miembros individuales. Cada uno debe juzgar, desde su actitud deliberativa personal, si las normas sociales y las prácticas que originan son aquellas que consideraría razonable acordar, si él y sus conciudadanos tuvieran la oportunidad de decidir conjuntamente los términos de su interacción. Más adelante consideraré un aspecto que me parece esencial para la concreción social de ese acuerdo. Pero lo que quiero subrayar ahora es, primero, que se parte de que la validez de las normas sociales depende y se deriva de normas aceptadas por los miembros individuales de la sociedad. No hay ninguna otra fuente reconocida por todos de donde puedan extraer su validez. Además, las normas sociales no deben derivar su validez de cualesquiera normas que los miembros individuales de la sociedad acepten (o de todas ellas), sino sólo de aquellas que todos los demás miembros pudieran aceptar también como válidas en las mismas circunstancias. La idea contractualista es que sólo las normas de la racionalidad deliberativa satisfacen esta condición. Todas las otras normas —derivadas, por ejemplo, de una religión, o de una concepción objetivista de los valores— son un tipo de normas que sólo los creyentes en la misma religión, o quienes comparten la misma cosmovisión, pueden aceptar como válidas para sí mismos. No se puede, por tanto, apelar directamente a este tipo de normas para defender prácticas sociales que deben ser justificables para todos.

## 2

Para entender más claramente la justificación contractualista, podemos contrastarla con una teoría moral y política bastante diferente, que también apela a la idea de un contrato social, y a la que, siguiendo a uno de sus defensores principales, me referiré como *s-contrac-*

1. Tomo la frase «tiene sentido» (*«makes sense»*) de Allan Gibbard, *Wise Choices, Apt Feelings*, Cambridge (Mass.), Cambridge University Press, 1990, pág. 7.

*tualismo*.[2] La diferencia entre estas dos concepciones fue oportunamente explicitada por Scanlon en su ya clásico artículo «*Contractualism and Utilitarianism*» (Contractualismo y utilitarismo).[3] Para formularla, propongo que nos tomemos ciertas libertades con el párrafo final de ese artículo, sustituyendo el interés en la protección personal por el interés más general de promover el bien propio, cualquiera que éste sea. El párrafo modificado queda como sigue:

> El contraste se puede explicar así. Desde un punto de vista, el interés en promover el propio bien es fundamental, y el acuerdo general se hace relevante como medio, o condición necesaria, para promover el propio bien. Desde el otro punto de vista, el s-contractualista, el deseo de promover el propio bien es un factor importante para determinar el contenido de la moralidad porque determina lo que puede ser acordado razonablemente. Pero la idea de un acuerdo general no surge como medio para avanzar el propio bien, sino que es algo más fundamental: aquello en lo que la moralidad consiste.[4]

El enfoque s-contractualista de Scanlon descubre la moralidad en la idea de un acuerdo general. La motivación de una persona para ajustarse a los principios y prácticas morales se deriva de «el deseo de poder justificar las propias acciones ante otros sobre premisas que no

2. Como se explica inmediatamente, el defensor de este punto de vista es T. Scanlon, y el término que emplea (y que Gauthier adopta) es «*contractualism*», frente al usado por Gauthier —y común en la tradición del contrato social liberal— de «*contractarianism*». Como es una diferencia imposible de mantener en español sin forzar mucho el idioma, he optado por traducir «*contractualism*» mediante el término «s-contractualismo», que aparecerá en lo sucesivo para distinguir las ocasiones en que Gauthier emplea «*contractualism*» (la teoría de Scanlon) en vez de «*contractarianism*» (su propia teoría) (*N. del t.*).

3. En *Utilitarianism and Beyond*, Amartya Sen y Bernard Williams (comps.), Cambridge, Cambridge University Press, 1982, págs. 103-128.

4. Véase «*Contractualism and Utilitarianism*», pág. 128. El pasaje original dice:

El contraste se puede explicar así. Desde un punto de vista, el interés en la protección es fundamental, y el acuerdo general se hace relevante como medio, o condición necesaria, para asegurar esta protección. Desde el otro punto de vista, el s-contractualista, el deseo de protección es un factor importante para determinar el contenido de la moralidad porque determina lo que puede ser acordado razonablemente. Pero la idea de un acuerdo general no surge como medio para asegurar la protección, sino que es algo más fundamental: aquello en lo que la moralidad consiste.

puedan rechazar razonablemente», dado que su motivación es similar.[5] Este deseo se satisface en la medida en que las propias acciones puedan considerarse de acuerdo con (o, al menos, no en contra de) los términos de un acuerdo general razonable. Una sociedad s-contractual permite un ámbito dentro del cual cada persona puede considerarse justificada ante sus semejantes.

Podríamos decir que, para Scanlon, cada persona se sitúa frente a todas las demás en una relación moral fundamental que se expresa en la justificación mutua. Habla de un deseo de poder justificar las propias acciones ante los demás. No cabe pensar que éste sea un deseo ordinario, como cualquier otro que la persona pueda tener. El deseo de poder justificar las propias acciones constituye el reconocimiento, por parte del agente, del significado y dignidad moral de sus semejantes. Un agente que no tuviera ese deseo, tampoco gozaría de este reconocimiento. Y para quien defendiera la idea de sociedad s-contractual propugnada por Scanlon, no reconocer la dignidad moral de otros sería, sin duda, una carencia moral.

Así, desde el punto de vista de Scanlon, la dirección de la justificación va del individuo a los otros. El papel del acuerdo es lograr que las acciones propias sean aceptadas por los otros. Sin embargo, en el contractualismo, que yo defiendo, la dirección de la justificación es precisamente la contraria. Partiendo de la idea de un deliberador racional, el contractualista no halla base alguna para postular la necesidad de justificar las acciones propias ante los otros. El papel del acuerdo es satisfacer la exigencia de cada persona de que las restricciones que la sociedad impone sobre ella estén justificadas; no un interés de justificarse a sí misma ante los demás.

Pero, ¿qué se logra mediante la justificación? Supongamos que una persona concede que ciertas prácticas sociales existentes superan el test contractualista. La tesis contractualista es que en ese caso la persona tendría una razón —normalmente una razón suficiente— para conformarse a esas prácticas. Aún más; tendría derecho a esperar que otros también se conformaran a ellas, y tendría una razón para exigir que aquellos que no lo hicieran fueran sancionados de algún modo por la sociedad. Y no tendría razón para quejarse si, por no haber cumplido con esas prácticas, ella misma fuera sancionada.

5. «*Contractualism and Utilitarianism*», pág. 116.

Obviamente, las sanciones también deberían superar el test contractualista. Pero si lo superasen, cada persona podría invocarlas justamente contra otros —y verlas justamente invocadas contra sí mismas— en caso de violación de las normas inherentes a las prácticas sociales. Esta doctrina de la justificación, que hace depender las razones en interacción del test de un acuerdo hipotético, es la esencia del contractualismo político.[6]

## 3

Podemos suponer que, en ausencia de restricciones sustantivas sobre la deliberación racional, cada deliberador procura actuar del modo que mejor realice sus valores. Así, como punto de partida para una investigación sobre las relaciones entre las pretensiones normativas de la sociedad y las actitudes normativas de sus miembros, podemos tomar la idea rawlsiana de que la sociedad es «una empresa cooperativa para el beneficio mutuo»,[7] entendida en el sentido de que cada miembro normal de la sociedad debe encontrar beneficioso pertenecer a ella, y a la vez debe hacer una contribución que los demás consideren valiosa. Quien no viera beneficio alguno en la sociedad, no tendría razones para aceptar las condiciones de interacción que impone. Quien no aportase nada útil para los demás, no les daría razón alguna para aceptarle dentro del ámbito de las condiciones sociales de la interacción. Por lo tanto, cada uno tiene que poder obtener de la sociedad algo que busca pero que no puede conseguir por sí mismo; y tiene que aportar a la sociedad algo que los demás buscan pero no podrían tener sin él. Dicho más sencillamente, cada uno tiene que ser a la vez beneficiario y benefactor.

Este requisito —de beneficiarse y a la vez beneficiar a los conciu-

6. No me ocupo aquí de defender esta doctrina. Sí defiendo una teoría de este tipo en *Morals by Agreement*, Oxford, Clarendon, 1986 (trad. cast.: *La moral por acuerdo*, Barcelona, Gedisa, 1994), cap. VI, y en «Assure and Threaten», *Ethics*, 104 (1994), págs. 690-721 (versión española, «Asegurar y amenazar», en este volumen) y en otros lugares. Pero una defensa adecuada requeriría situar la doctrina dentro de una teoría general de la deliberación racional.

7. John Rawls, *A Theory of Justice*, Cambridge (Mass.), Harvard University Press, 1971, pág. 4.

dadanos— tiene como función restringir el tipo de valores, o proyectos vitales, que los miembros individuales de la sociedad pretenden realizar. Tal y como estoy describiendo el contractualismo, podría dar la sensación de que en el proceso de lograr un acuerdo sobre las instituciones y prácticas sociales, y sobre las obligaciones que imponen, cada uno debería deliberar desde una concepción o proyecto personal de vida satisfactoria determinado con anterioridad, sea cual sea fuere concepción o proyecto. Las instituciones sociales aceptables serían entonces aquellas que se amoldaran a los proyectos vitales de los miembros de una sociedad, quienes esperarían que su participación en el producto social conjunto ayudara a la realización de su proyecto. De hecho, la sociedad se ajustaría a los proyectos vitales de un modo similar a como el mercado perfectamente competitivo se ajusta a las demandas del consumidor para satisfacerlas.[8]

Ahora bien, existen varias objeciones a este modo de entender el argumento contractualista. La más profunda, a la que volveré más adelante, es que reduciría las normas sociales al papel instrumental que, como he dicho, es sólo una parte de sus funciones —en la medida en que no sólo establecen limitaciones sino que a la vez crean oportunidades—. Pero además debemos señalar que no todos los proyectos vitales, o concepciones de una vida valiosa, posibles dentro de una estructura social dada, convierten a sus poseedores en contribuyentes netos al producto conjunto que la interacción social hace posible. La sociedad no puede, desde un punto de vista racional, aceptar a nadie que elija un proyecto de vida que imponga costes netos sobre sus conciudadanos; este individuo sería, más que un benefactor, un *malefactor*. Imponer costes a los demás violaría lo que he identificado como el requisito fundamental en que se basa la interacción racional voluntaria.[9] Este requisito, que generaliza la interpretación de la salvaguardia lockeana empleada por Nozick, prohíbe las interacciones en las que una parte gana y otra pierde en relación a la situación inicial de no interacción. Al considerar la sociedad como una empresa cooperativa

8. Defiendo este concepto de una «sociedad esencialmente justa» en *Morals by Agreement*, págs. 340-341. Sin embargo, argumentaré que no parece ser una parte necesaria —ni factible— de una teoría contractualista.

9. *Véase Morals by Agreement*, págs. 200 y sigs.; también mi artículo «Uniting separate persons», en *Rationality, Justice and the Social Contract*, David Gauthier y R. Sugden (comps.), Hemel Hempstead, Harvester-Wheatsheaf, 1993, págs. 182-185.

para el beneficio mutuo, excluimos las violaciones graves de la salvaguardia, puesto que incluso nuestra idea de lo que es una interacción social permisible queda condicionada por ella.

Ilustraré la aplicación de esta salvaguardia con un ejemplo deliberadamente provocativo. Consideremos a las mujeres que, teniendo como único ingreso un subsidio estatal, deciden tener un hijo o, al quedar embarazadas, deciden no abortar. Las mujeres que así actúan han adquirido hoy en día una mala imagen que, desde el punto de vista contractualista, tendrían bien merecida. Al elegir traer un niño al mundo a pesar de su carencia de los recursos financieros y, muchas veces, emocionales, para cuidarlo adecuadamente, están violando una extensión de la salvaguardia que prohibiría interacciones que mejoraran a una parte haciendo que otra quede peor de lo que nadie tendría por qué estar. Hablo de una extensión de la salvaguardia porque en este caso la parte perjudicada —el niño— llega a existir sólo gracias a la interacción, de modo que no se puede comparar la situación del niño con aquella en que estaría al margen de la interacción. Sin embargo, se puede decir que la situación del niño implica costes que no sería necesario que nadie soportase, y que nadie soportaría si las mujeres que carecen de recursos financieros y emocionales para dar a sus hijos el cuidado necesario no los tuvieran.

Se puede objetar que la violación de la salvaguardia no es aquí imputable a las madres con pocos recursos, o a los padres, frecuentemente ausentes; sino más bien a la sociedad, que no facilita los medios necesarios para cuidar de sus hijos. Pero, al menos en principio, esta objeción no es concluyente. Porque si la sociedad tuviera que proveer esos medios, ello representaría un coste neto para otros miembros de la sociedad. Dicho de otra forma, la madre subsidiada estaría imponiendo un coste neto sobre sus conciudadanos por elegir un cierto proyecto vital. Su contribución bruta al producto social conjunto es menor que el coste bruto que otros miembros de la sociedad han de soportar para proveer una atención adecuada a su hijo. Así, aunque no violara la salvaguardia por el hecho de tener el hijo, lo haría al imponer el coste de su crianza sobre sus conciudadanos.

Para evitar malentendidos, hay que advertir que la violación de la salvaguardia no se debe al hecho de que una mujer tenga un subsidio y a la vez sea madre. Es evidente que cualquier persona podría encontrarse en esa situación debido a la mala suerte, en contra de toda

expectativa razonable. La violación se produce porque una mujer que vive de un subsidio decide tener un hijo, o porque una mujer decide tener un hijo sabiendo que va a necesitar una subvención para criarlo. Todo el mundo está sujeto al riesgo de llegar a ser incapaz de proveer para sí mismo (directamente o mediante el intercambio). La sociedad proporciona una seguridad contra este riesgo. Pero no por ello autoriza a las personas a que elijan no proveer para sí, o a tener hijos sin ser capaces de proveer para ellos. No autoriza a las personas a que elijan ser un coste neto para sus semejantes, o que elijan traer a la existencia a aquellos que lo van a ser.[10]

La salvaguardia no es un requisito utilitarista. Cada miembro de la sociedad es libre para elegir un proyecto vital que responda a sus intereses, sin tener en cuenta el bienestar general de la sociedad. Pero

10. Hay otra objeción, más difícil, contra la tesis de que las mujeres subsidiadas que deciden tener un hijo violan la salvaguardia. Quiero mencionar esta objeción, aunque no puedo examinarla en detalle. La objeción es que nuestra sociedad impide que las personas que se encuentran en situaciones como la de una madre subsidiada tengan la oportunidad de hacer contribuciones útiles a la sociedad, sobre todo si son solteras; a pesar de no existir ninguna razón objetiva para ello. Si éste fuera el caso, entonces la injusticia que la madre parece estar perpetrando contra el hijo y/o la sociedad como un todo se debería sólo, o al menos en gran parte, a que la sociedad no le ha dado las suficientes oportunidades, lo cual parece una grave injusticia social. Típicamente las madres subsidiadas son excluidas de trabajos útiles que les podrían dar los medios para criar adecuadamente a sus hijos. Pero para ver más claramente la naturaleza y la fuerza de esta objeción quizá sea interesante plantear el problema general creado por el desempleo estructural.

Supongamos, de acuerdo a la opinión de la mayoría de economistas, que una sociedad capitalista eficiente requiere un cierto nivel no trivial de desempleo. Parece obvio que, como consecuencia, algunas personas estarán desempleadas sin que pueda considerarse culpa suya. El producto social global disminuiría si estas personas se sumasen a la población activa. Y, por tanto, su presencia en la sociedad no es un coste neto para los demás. Si, para maximizar el producto social global es necesario cierto nivel de desempleo, entonces resulta que los desempleados están, paradójicamente, contribuyendo al beneficio de todos.

Los desempleados por razones estructurales no tienen por qué sentirse tratados injustamente. Excepto si su desempleo se toma como excusa para negarles una parte razonable del producto social global. Si la situación de las madres subsidiadas fuera ésta, entonces la segunda objeción sería plausible. Y se da la posibilidad adicional de que las madres subsidiadas —de nuevo, especialmente si son solteras— carezcan de oportunidades, no por razones estructurales válidas, sino por una discriminación directa que ignora su contribución potencial al producto social global.

Como advertí, no intentaré evaluar aquí estos problemas.

cualquier comprensión plausible de la tesis normativa de que la sociedad es una empresa cooperativa para el beneficio mutuo —una tesis que subyace a la posibilidad de ofrecer una justificación contractualista de la organización social— excluirá algunos proyectos vitales concebibles. En ninguna sociedad cuyas normas hayan de superar el test contractualista puede darse una libertad ilimitada para formar y ejecutar cualquier proyecto vital.

4

Al hablar de una «libertad para formar», estoy suponiendo que el fin que una persona valora y trata de alcanzar, al cifrar en él el significado de una vida valiosa, está en alguna medida bajo su control y elección. Esto proporciona un claro fundamento para rechazar la idea de que las normas sociales acordadas deben ser estrictamente neutrales entre los proyectos vitales de los miembros de la sociedad, adaptándose siempre a ellos, sean los que fueren. Suponer que las normas sociales acordadas *tienen que* exhibir este tipo de neutralidad entre planes de vida posibles como único modo de asegurar su justificación ante los individuos, significa tratar esos planes como dados, no sólo desde el punto de vista de la sociedad, sino desde el de cada uno de sus miembros individuales. Si cada uno de nosotros se encontrara con un conjunto de fines que definieran una vida valiosa para él, y que fuera incapaz de abandonar o modificar sin un grave daño para sus perspectivas de lograr una vida feliz, entonces una sociedad contractualista tendría, ciertamente, que ser neutral respecto a estos planes de vida basados en fines dados; y tendría que intentar ajustarse de modo que cualquier proyecto, fuera el que fuere, recibiera una expectativa de satisfacción que reflejase imparcialmente la contribución realizada por su propietario al producto conjunto de la sociedad. Pero los seres humanos no son receptores pasivos de un conjunto de fines o un proyecto vital. Al contrario, los valores, fines y proyectos están sujetos a la deliberación y la elección.

Además, la misma idea de que las normas sociales sean neutrales entre proyectos vitales es incompatible con el reconocimiento de que dan lugar a prácticas que amplían el horizonte de fines, y no simplemente el horizonte de medios. Así, los proyectos vitales que un agen-

te puede adoptar dependen de las prácticas de su sociedad. No cabe formularlos independientemente, y usarlos luego como base para seleccionar una estructura social, incluso aunque la adecuación entre los proyectos y la estructura social se conciba dinámicamente empleando el modelo del mercado. De hecho, el propio modelo del mercado no capta el papel que el mercado juega en la ampliación del horizonte de bienes disponibles. En conclusión, el contractualista no supone que las prácticas sociales tengan que justificarse mostrando que satisfacen proyectos vitales independientemente dados, sino mostrando, más bien, cómo es que permiten y fomentan la formación de proyectos vitales que pueden ser satisfactoriamente realizados en sociedad.

Supongamos que cada miembro de la sociedad es definido por un proyecto vital. Supongamos además que cada persona ve su proyecto como algo que ella elige, una opción y no una imposición, de modo que puede imaginar que la elección de un proyecto completamente diferente sería compatible con su integridad como persona. Más adelante necesitaré cualificar esta suposición. Pero comenzaré admitiendo que la que llamaré actitud normativa de cada persona se reduce a su naturaleza de deliberador racional, de modo que una sociedad contractualista queda adecuadamente justificada desde la actitud normativa de cada uno de sus miembros si está constituida de acuerdo a un conjunto de normas que pudiera ser acordado por deliberadores racionales como expresión de las condiciones de su interacción.

Con esto, puedo describir de modo muy abstracto cómo se resolvería (al menos en parte) el ajuste entre los individuos y la sociedad impuesto por el acuerdo entre deliberadores racionales sobre los términos de su interacción, dado que todos reconocen que esos términos crean posibilidades de nuevos fines, y no simplemente nuevos medios para fines ya existentes.[11] La sociedad debe estar estructurada de tal modo que, primero, cada persona normal disponga efectivamente de una serie de posiciones sociales, cada una de las cuales demande una contribución productiva de quien la ocupe, y le recompense de acuer-

11. Lo que sigue puede leerse como una interpretación del principio del beneficio relativo *maximin* (véase *Morals by Agreement*, págs. 154-155), adaptada al acuerdo político sobre los términos de la interacción.

do a su contribución marginal al producto social global. Segundo, cada una de estas posiciones debe ser compatible con un conjunto de proyectos vitales, cualquiera de los cuales podría, en circunstancias normales, ser elegido y realizado satisfactoriamente por el ocupante de esa posición, dado que contaría con una participación en el producto social que recompensaría el desempeño efectivo de la posición. Para poder decir que un individuo dispone realmente de una posición, ésta debe ser adecuada a sus capacidades, que el individuo debe haber tenido la oportunidad de desarrollar y ejercitar. Y para que la realización de los proyectos del individuo sea satisfactoria, la posición debe adecuarse a sus motivaciones y afectos, que debe haber tenido la oportunidad de educar. Más adelante volveré a la cuestión de cómo asegurar el acceso a la preparación para la elección de una posición social y un proyecto vital.

Nótese que esta descripción del ajuste entre los individuos y la sociedad se basa principalmente en el grado de realización de un proyecto vital satisfactorio, y sólo de modo secundario en cómo quedan unos individuos en relación con otros. Si cada persona puede elegir entre las distintas posiciones productivas, cada una con su recompensa apropiada, y puede luego elegir entre proyectos vitales a los que esa recompensa le permite dedicarse satisfactoriamente, entonces —mientras estas elecciones se ofrezcan a todos por igual— nadie tiene razón para considerar que su posición en la sociedad, o sus relaciones con sus conciudadanos, son gravemente injustas. El punto fundamental es que haya suficientes oportunidades; mientras que, dadas las diferencias naturales entre las personas, hay que reconocer que la igualdad en el resultado no es ni posible ni deseable.

Esta descripción no pretende ser completa; en concreto, omite las libertades y derechos políticos fundamentales que permitirían a cada individuo incluir en su proyecto vital la participación en el gobierno democrático de la sociedad, así como las obligaciones que garantizasen un nivel mínimo de participación. Son suficientemente conocidos y no serán objeto del presente estudio.

A veces se supone que una sociedad contractualista debe caracterizarse por que absolutamente todas las relaciones entre sus miembros sean contractuales. Sin embargo, ningún elemento de la descripción que acabo de exponer lo exige, ni —bien entendido— lo sugiere. Una cosa es suponer que las personas deben considerar los

términos de su interacción como el resultado de un acuerdo; y otra muy diferente es suponer que las personas deban creer que el resultado del acuerdo exige que todas las interacciones se establezcan sobre una base contractual, o que cualquier interacción no contractual tenga que ser siempre voluntaria (no obligatoria). De hecho parece evidente que en cualquier sociedad habrá algunas relaciones entre personas que no son directamente elegidas o acordadas por algunas de las partes implicadas en ellas y que, sin embargo, imponen deberes y otras restricciones. El contractualista no tiene por qué rechazar la institución de la familia —pese a que no pueda concebirse como el resultado directo de un acuerdo entre todos sus miembros— en la medida en que la vida familiar pueda concebirse como algo tan esencial para una sociedad, que sería un producto posible de un acuerdo universal razonable.

## 5

Quizá la idea de un acuerdo entre deliberadores hipotéticos que eligen los términos de su interacción parezca demasiado vaga para proporcionar conclusiones directas sobre la justificabilidad de normas y prácticas sociales reales. Después de todo, no hay ningún sistema social que todos consideren el mejor. ¿No ocurrirá siempre que cualquier norma propuesta será rechazada por algunos?

El error es esperar que el resultado del contrato social, o de cualquier otro acuerdo, sea el mejor posible para todas las partes. Lo que importa es que hay estructuras sociales que son sin duda universalmente preferibles al vacío que existiría en ausencia de normas y prácticas, de modo que cada parte posee un gran incentivo para ponerse de acuerdo con las demás. E importa también que la estructura social real no sólo sea mutuamente beneficiosa, sino que además asegure que la proporción entre los beneficios y la contribución que cada uno decide aportar sea aproximadamente la misma para todos. Los términos de una empresa cooperativa son los mismos para todos en cuanto asignan el beneficio en proporción a la contribución de cada uno, equitativamente. Y si las personas tienen un gran incentivo para llegar a un acuerdo con las demás, nadie encontraría razonable dificultar o arriesgar la posibilidad de ser

aceptado en un acuerdo por exigir un trato más favorable que los demás.

Así, la posibilidad de que otras normas, prácticas e instituciones pudieran ser igualmente razonables no es una razón contra el orden social existente. Sólo si la sociedad existente no pudiera concebirse plausiblemente como el resultado de un acuerdo razonable sobre los términos de la interacción habría una base para rechazarla, o para tratar de reemplazarla por normas y prácticas que pudieran concebirse como el resultado de un acuerdo. La tarea de la justificación contractualista queda suficientemente determinada mediante la exclusiva apelación a la racionalidad deliberativa de los miembros de la sociedad.

## 6

Al comienzo caractericé el contractualismo como individualista, minimalista y constructivista. Ahora quiero añadir algo sobre estos rasgos, lo cual me llevará a discutir el lugar que ocupan las normas no contractuales en el marco de una sociedad contractualista. En primer lugar, el individualismo que la justificación contractualista expresa es bastante débil. Aunque la justificación debe dirigirse a los individuos, concebidos como deliberadores racionales, las prácticas así justificadas no tienen por qué ser individualistas. Ya he comentado que las relaciones no contractuales entre personas son susceptibles de justificación contractualista, y aquí quiero simplemente generalizar esta idea.

La sociedad humana comprende prácticas e instituciones que poseen estructuras normativas propias. El test contractualista no impone ningún límite formal a las características de esas estructuras. Los derechos de grupos y las responsabilidades colectivas son susceptibles de ser sometidos al examen contractualista, aunque es dudoso que lo superasen. Pero en principio, se podría desarrollar una versión comunitarista o incluso colectivista del contrato social, igual que John Harsanyi ha desarrollado una versión utilitarista.[12]

12. Véase especialmente «Morality and the Theory of Rational Behavior», *Social Research*, 44 (1977), págs. 625-656.

No es que yo diga que comunitaristas o colectivistas vayan a alegrarse de la posibilidad de una defensa contractualista de sus ideas. Como he dicho, dudo que sea probable una defensa de este tipo. Además, si pudiera darse una verdadera defensa, mostraría únicamente que la adhesión a las normas comunitarias promovería el beneficio mutuo de deliberadores racionales, y éste no es el tipo de justificación que buscan quienes se consideran comunitaristas. Ellos no aceptarían la prioridad del deliberador individual en la determinación de la validez normativa de la comunidad. Es improbable que aceptaran la tesis de que lo que puedan tener de bueno los derechos de grupos, las responsabilidades colectivas y otras ideas similares, deba revelarse mostrando que las practicas e instituciones que los incorporan superarían el test contractualista.

Pero si esto es así, ¿qué interés tiene mencionar, e incluso subrayar, la posibilidad de que normas no individualistas pudieran superar el test contractualista? Lo que pretendo es aclarar que la relación entre la racionalidad deliberativa y otras concepciones normativas es, por así decir, justificatoria más que definitoria. Lo cual afecta a la comprensión del minimalismo y constructivismo de mi explicación, pues ambos rasgos deben entenderse en sentido justificatorio. Para defender una práctica o institución social sólo podemos apelar a una justificación que la haga depender de la racionalidad deliberativa de los miembros de la sociedad. Y tal justificación consiste en mostrar que tendría sentido para los miembros de una sociedad acordar una estructura social que incluyera esa práctica o institución. Pero no es necesario que los elementos de esa estructura estén enteramente compuestos por (o construidos a base de) las normas de la racionalidad deliberativa. Aunque la sociedad pueda ser conceptualizada como una construcción realizada por individuos deliberativamente racionales, los elementos de esa construcción no tienen por qué reducirse necesariamente a los recursos normativos de la racionalidad deliberativa.

En trabajos anteriores, como en mi artículo «Public Reason»[13] (Razón Pública), no distingo entre los papeles justificatorio y definitorio de la racionalidad deliberativa, lo cual me lleva a aceptar una forma más estricta de constructivismo. Y todavía estoy convencido de que gran parte de los recursos normativos empleados para la cons-

13. En *Social Philosophy and Policy*, 12 (1995), págs. 19-42.

trucción de las instituciones sociales, especialmente la autoridad social, provienen directamente de la normatividad deliberativa de los individuos. Pero en este momento creo que la idea de que la autoridad social *tiene que* construirse a base de las normas deliberativas individuales, confunde la necesidad de justificar la autoridad en términos de racionalidad deliberativa, con el intento de crear una autoridad transfiriendo a un agente social la racionalidad deliberativa de los miembros individuales de la sociedad.[14]

## 7

Hasta este punto he estado suponiendo que la actitud normativa relevante de los miembros individuales de la sociedad se reduce a su capacidad como deliberadores racionales. Ahora bien, hay dos modos de entender este supuesto. El primero es verlo como parte de lo que podríamos llamar una política agnóstica. Desde ese punto de vista, las normas que no puedan constituirse o justificarse en términos de deliberación racional no son públicamente afirmadas ni negadas; simplemente se consideran inadecuadas como objeto de la justificación social, por no poder conformar una base común para todos los miembros de la sociedad. Tales normas se tratan públicamente como incognoscibles, aunque pueden no serlo desde el punto de vista de los individuos. El otro modo de entender el supuesto es, diríamos, ateo: niega la validez de las normas que no puedan constituirse o justificarse mediante la deliberación racional.

En una sociedad donde las personas no sólo tienen ideas sustantivas diferentes de lo que es bueno y correcto, sino también ideas justificatorias diferentes sobre el fundamento de lo bueno y lo correcto, sólo la interpretación agnóstica es factible o permisible. Algunas personas están comprometidas personalmente con la idea de ciertos valores objetivos y de una moralidad independiente de la deliberación —de base secular o religiosa— y, para ellas, aceptar estructuras sociales opuestas a sus creencias comprometería su integridad como personas. La comprensión de este hecho prohíbe la negación de nor-

14. Entiendo que éste es el objetivo central del proyecto de Thomas Hobbes en el *Leviathan*, y el mío en «Public Reason».

mas no pertenecientes a la deliberación o no justificadas por ella, pero a cambio hace necesaria la demanda agnóstica de que tales normas no sean usadas en la justificación social. Porque aceptar estructuras sociales basadas en concepciones normativas no compartidas podría comprometer la integridad de las personas que no las compartieran. Muchos rechazarán que la normatividad de la racionalidad deliberativa sea suficiente; pero no se puede negar que ofrece una base común para quienes estén dispuestos a aceptarla. Cada cual puede tener diferentes ideas sustantivas y justificatorias sobre el bien y lo correcto. Pero quienes, con base en sus particulares convicciones, rechacen la racionalidad deliberativa, carecen de una base común sobre la que edificar una unión con quienes no comparten su punto de vista normativo global, y no pueden ser incluidos en una sociedad constituida como una «empresa cooperativa para el beneficio mutuo».

Volvamos a la idea de que cada miembro de la sociedad queda definido por un proyecto vital. Sobre este punto, hemos de precisar ahora que, aunque algunos elementos del proyecto personal de vida pueden tratarse como opcionales, otros pueden ser percibidos como obligatorios. Los cristianos no suponen que creer en un dios o las obligaciones del culto sean aspectos opcionales de su plan de vida. Los kantianos no suponen que adecuar las máximas de sus acciones al imperativo categórico sea opcional; los utilitaristas no suponen que sea opcional dirigir sus acciones hacia la mayor felicidad del mayor número. Se trata de obligaciones —diferentes para diferentes personas— que reflejan compromisos no incluidos en la idea de la racionalidad deliberativa.

Una sociedad contractualista debe ser públicamente agnóstica respecto a estos compromisos. Pero debe serlo de un modo benevolente, en la medida en que quienes mantienen estas creencias compartan el compromiso de interactuar con quienes profesan diferentes convicciones en términos que para todos tenga sentido aceptar. No hay que privilegiar ninguna perspectiva, pero tampoco hay que prohibir o poner trabas a ninguna. Y mi interpretación es que ésta es la intención de las cláusulas sobre la religión de la Primera Enmienda, que podemos ampliar para cubrir cualesquiera convicciones evaluativas y normativas.

Para ilustrar las implicaciones de un agnosticismo benevolente,

consideremos cómo podría una sociedad contractualista aceptar a los sabatarios estrictos —personas que creen que un día universal de descanso es una obligación religiosa—. Por un lado, son libres, desde luego, para defender que se imponga como práctica social un día universal de descanso; pero desde la perspectiva pública, esta postura sólo puede ser tratada como su preferencia particular, que habrá de ser contrastada con otras preferencias, sin privilegiarla por el hecho de tener un fundamento pretendidamente religioso. Un día universal de descanso limitaría ciertos proyectos vitales, pero facilitaría otros. Hasta donde soy capaz de juzgar, no sería claramente incompatible con la garantía de que cada persona disfrutase de un amplio abanico de proyectos vitales satisfactorios entre los que elegir y, por lo tanto, parece una opción admisible para una sociedad contractualista, aunque quizá no muy deseable, según mi opinión.

Por otro lado, los sabatarios tienen una obligación que acomodar en sus propios proyectos vitales; una obligación que, para ellos, va más allá que cualquier pretensión nacida de una preferencia particular. La reclamación de libertad para descansar en el *Sabat* no representa, desde su punto de vista, un fragmento opcional de su proyecto vital, sino un deber. Hemos de suponer que, si las personas estuvieran en condiciones de elegir los términos de su interacción, se cuidarían mucho, sin duda, de darse libertad para elegir proyectos vitales que les permitieran vivir conforme a las obligaciones morales o religiosas en que creyesen; al menos mientras tales proyectos vitales no les impidiesen ser contribuyentes netos al producto social global. Así, en una sociedad contractualista se daría la posibilidad de que los sabatarios eligieran una vida conforme a sus convicciones; aunque ello les costaría no tener tantas posibilidades como otros miembros de la sociedad —por ejemplo, no podrían ser deportistas profesionales.

Hay que subrayar que este argumento, que justificaría medidas de protección para las convicciones morales y religiosas de los individuos (medidas no aplicables a sus simples preferencias), no apela al contenido de esas convicciones, sino sólo a la actitud que esos individuos tendrían hacia la moralidad y la religión. Las prácticas e instituciones de una sociedad contractualista deben estar justificadas ante personas concebidas como deliberadores racionales, que imaginan la posibilidad de determinar las condiciones de su interacción. He supuesto que, ante este problema, la preocupación principal de cada

persona sería garantizar que dispone de un número suficiente de posiciones sociales, cuyo desempeño le permitiera desarrollar un cierto número de proyectos vitales capaces de proporcionarle, a su vez, una vida satisfactoria (y compatible con la posición). Pero en la medida en que la elección entre proyectos vitales puede estar condicionada por obligaciones morales o religiosas, las personas que profesen tales creencias y tengan sus preceptos por obligatorios querrán garantizar, además, que las posiciones y proyectos factibles incluyan opciones que les permitan cumplir esas obligaciones. Evidentemente, el requisito fundamental del contractualismo, el beneficio mutuo, limita el alcance de esa garantía. Pero, aun sin compartir las convicciones morales o políticas concretas de nuestros conciudadanos, podemos entender que todos procuren, como deliberadores racionales, obtener tal garantía.

## 8

Para poner otro ejemplo de las implicaciones del agnosticismo contractualista, pensemos en dos prácticas de caza: la caza del zorro, tal como la practican algunos nobles ingleses, y la caza de focas, tal como se practicaba tradicionalmente en las comunidades de la costa atlántica canadiense. Algunas personas consideran crueles estas prácticas, y preferirían pertenecer a una comunidad donde estuvieran prohibidas por la ley. En una sociedad contractualista, tienen derecho a defender públicamente esta preferencia. Pero debemos preguntarnos qué efectos tendría la ilegalización de estas prácticas sobre las oportunidades productivas y las elecciones de proyectos vitales de las personas que ahora participan en ellas, y qué tratamiento estarían recibiendo sus intereses en comparación con los de sus conciudadanos.

Los cazadores del zorro, por mucho que disfruten de este deporte, no verían disminuida su contribución social y, una vez abolida la caza, encontrarían otras opciones recreativas disponibles. Sin duda que tendrían que hacer cambios en sus planes vitales, pero no se podría alegar plausiblemente que estos cambios disminuyesen su calidad de vida. Así, la legitimidad de la caza del zorro sería decidible por los cuerpos legislativo y regulador ordinarios de una sociedad contractualista. El agnosticismo en la comparación entre los valores

de los cazadores del zorro y sus oponentes lleva a ver públicamente esos valores como preferencias opuestas, que no implican ninguna cuestión pública de principio.

Por su parte, los cazadores de focas se verían mucho más afectados por una prohibición —tanto en sus actividades productivas como en su proyecto general de vida— y, teniendo en cuenta el ya de por sí limitado abanico de posibilidades efectivas de que disponen, una disminución adicional no compensada supondría un perjuicio muy superior al de los demás miembros de la sociedad. Evidentemente, la forma de vida de los cazadores de focas podría llegar a hacerse inviable si las personas que consideran cruel esa caza deciden, con toda justificación, negarse a comercializar o comprar sus productos. Pero ésa es otra cuestión. No es legítimo que una sociedad contractualista cierre áreas de elección que han llegado a hacerse fundamentales para algunos de sus miembros, a no ser que mantener esas áreas haga que la presencia de esos miembros suponga un coste neto a la sociedad. Y ésta no es una cuestión que pueda decidirse mediante los procedimientos legislativo y regulador ordinarios. Afecta a términos fundamentales del contrato social.

Quienes consideran cruel la caza de focas siempre pueden apelar al coste de sus sentimientos. Pero éstos están en gran medida bajo su control. Pueden elegir actividades satisfactorias que no les lleven a estar siempre atormentados por la muerte a garrotazos de las crías de foca. No tienen más motivo de queja (por la disminución en calidad de vida) que aquellos a quienes se niega el placer de la caza del zorro. Sin duda que a muchas personas les parecerá que este razonamiento es totalmente desatinado. Porque la oposición a la caza de focas no se considera como parte de un conjunto de valores que uno elige, pero podría no elegir. Quienes se oponen, no creen que su punto de vista sea optativo, sino una obligación. Esa perspectiva quizá justifique que se nieguen por completo a participar en actividades relacionadas con la caza, y que la sociedad les proporcione oportunidades compatibles con esa abstención; pero no puede justificar su pretensión de restringir las elecciones de quienes no comparten sus convicciones. El agnosticismo de una sociedad contractualista exige que sus convicciones se acepten como meras preferencias. Y la satisfacción de su preferencia de prohibir la caza de focas sería inadmisible, porque limitaría seriamente la elección de posiciones y proyectos vitales satisfactorios para otros miembros de la sociedad.

## 9

El agnosticismo benevolente que la sociedad contractualista aplica a las diferentes actitudes morales y religiosas no debe llevarnos a suponer que la sociedad es ella misma moralmente agnóstica. Esto debería deducirse claramente de mi insistencia en la obligación del beneficio mutuo, al considerar la provisión de una garantía para el ejercicio individual de las convicciones morales y religiosas. El agnosticismo contractualista se limita a aquellos puntos de vista que no son ni validados ni rechazados en términos de racionalidad deliberativa. Pero el contrato social sí tiene esa validación, y una sociedad contractualista espera y exige la adhesión de sus miembros a él. Si los miembros de una sociedad contractualista alegan sus particulares concepciones evaluativas, religiosas, o morales contra las prácticas e instituciones constituidas conforme a los términos del acuerdo, se están excluyendo de la estructura de la cooperación social.

Creo que éste es el verdadero significado de la defensa rousseauniana de la religión civil.[15] Es cierto que algunas de las provisiones particulares que Rousseau supone necesarias para la existencia estable de la sociedad, a nosotros ya no nos lo parecen, pero ello no desacredita la idea de una moralidad públicamente reconocida por todos los miembros de la sociedad y que todos están obligados a cumplir. Es notable que Rousseau incluye en su religión civil el rechazo de la intolerancia, de la idea de que la sociedad reconoce un único camino para la salvación; esto se corresponde con el agnosticismo benevolente de una sociedad contractualista. Por otro lado, su propuesta de castigar con la muerte a quienes se hubieran comprometido con la religión civil y luego la negasen, nos parece hoy de todo punto excesiva. Pero desde luego, no hay ninguna razón contractualista que exija proveer de algún modo para las vidas o proyectos vitales de quienes, teniendo la oportunidad de vivir conforme a términos que les harían contribuir al producto social conjunto, eligen, por cualquier motivo, hacer su presencia costosa para sus conciudadanos. Incluso aunque su exclusión refleje sus convicciones morales o religiosas, no pueden apelar al agnosticismo benevolente de una sociedad contractualista, pues éste sólo alcanza a quienes aceptan su lugar

15. Jean-Jacques Rousseau, *Contrat social*, libro IV, cap. 8.

como socios en una empresa cooperativa en condiciones que tendría sentido para todos aceptar, en cuanto agentes deliberativamente racionales.

## 10

En un punto anterior mencioné la importancia que tiene para cada individuo que se asegure su acceso a la preparación necesaria para la elección de una posición social y un proyecto vital; preparación que incluye tanto el desarrollo y formación de sus capacidades como la educación de sus motivaciones y afectos. Los niños, como tales, son miembros improductivos de la sociedad; sin embargo, sus derechos e intereses deben ser reconocidos en el contrato social. Las partes del contrato son seres humanos individuales que, al acordar los términos de la interacción social, tienen en cuenta su vida entera y, por tanto, tienen interés en asegurar, en la mayor medida posible, que van a disfrutar durante su niñez de seguridad económica y emocional, y que van a disponer de formación para sus capacidades y educación para sus afectos de modo que puedan integrarse en una vida adulta satisfactoria y productiva. El requisito contractualista de que cada uno debe esperar no ser sólo beneficiario sino también benefactor no se aplica a cada momento de la vida de un individuo, sino al curso de su vida como un todo.

Asegurar una buena educación a cada miembro de la sociedad impondrá serias obligaciones sobre el Estado y los padres; estas obligaciones superarán con mucho las obligaciones mínimas impuestas por la salvaguardia. Ninguna sociedad que dé a todos sus miembros la seguridad de una buena educación puede reconocer que las personas tengan derecho a procrear más o menos todo lo que quieran. Si una infancia económica y emocionalmente segura es una pretensión básica consagrada por el contrato social, los futuros padres habrán de aceptar las condiciones para la realización de esa pretensión. Creo que podemos plausiblemente suponer que, en la hipótesis de tener que acordar las condiciones de la interacción, nadie estaría dispuesto a correr el riesgo de crecer en condiciones de indigencia emocional o material, sólo para ganar la libertad de tener hijos, o de concebirlos irresponsablemente, de modo que tuvieran que criarse en las mismas circunstancias.

Sin duda que en una sociedad contractualista tecnológicamente avanzada, la pobreza sólo afectaría a las personas que no aprovecharan las oportunidades para elegir un proyecto vital productivo y satisfactorio. Esa oportunidad no se negaría a nadie, como sí se niega a muchas personas en nuestra sociedad. Pero un proyecto vital que puede ser muy apropiado para una persona soltera, puede no serlo para quien desea tener hijos. Si, como parece probado desde un punto de vista psicológico, la madurez social de un niño mejora significativamente cuando su educación se desenvuelve en un entorno emocionalmente afectuoso y estable en el cual el niño desarrolla fuertes lazos afectivos con un adulto de cada sexo —especialmente si se trata de sus padres naturales—, entonces las personas que deseen tener hijos tendrán que comprometerse a crear y mantener en lo posible tal entorno. En modo alguno se trata de condenar a quienes no quieran, o no puedan, hacerlo. Se trata simplemente de indicar el tipo de restricciones que la demanda de una educación eficaz (que ha de formar parte del conjunto de los términos de la interacción aceptables para todos) impone sobre los proyectos vitales aceptables, cuando éstos afectan a niños.

No puede haber ninguna garantía contra la mala suerte ni, ay, la malevolencia. En estos casos, serán necesarios arreglos subóptimos para criar a los niños. Estos arreglos pueden incluir la paternidad de individuos solteros cualificados, o de parejas homosexuales, así como la adopción por parte de parejas heterosexuales. Quizá ninguna de estas posibilidades sea la más deseable para el desarrollo psicológico del niño, pero ello no significa que no sea preferible a las alternativas disponibles en los casos en que uno de los progenitores ha muerto, o queda incapacitado, o cuando se demuestra que uno, o los dos, son gravemente incompetentes o peligrosos. Si hay que elegir entre una crianza que va a producir la identificación con los padres naturales y una crianza que ofrezca afecto y estabilidad, hay que preferir sin duda esto último. Pero será una opción impuesta por las circunstancias. Creo que la organización reproductiva y familiar de una sociedad contractualista tenderá a que sean los padres naturales quienes se encarguen conjuntamente de la crianza de cada niño, y la justificación de las restricciones que este objetivo impondrá a los proyectos vitales individuales de madres y padres puede hallarse en que las partes del contrato social no están dispuestas arriesgarse a

sufrir una infancia en condiciones de indigencia emocional o material.

Las mencionadas limitaciones en los proyectos vitales no implican ninguna evaluación negativa de la vida en solitario, como miembro de una pareja homosexual o, para el caso, como miembro de una comuna, en relación con la vida en una pareja heterosexual; excepto en la medida en que estas opciones tienen efectos diferenciales sobre el desarrollo del niño. Que tengan estos efectos y, en concreto, que *ceteris paribus* los lazos afectivos entre un niño y una pareja heterosexual de adultos favorezcan su desarrollo emocional son afirmaciones de hecho que, si bien parecen plausibles, no podrían ser establecidas mediante la discusión filosófica. Pero, al menos desde mi punto de vista, la filosofía política tiene que emplear suposiciones sobre el desarrollo humano que no puede establecer. Sólo así puede mostrar alguna conexión entre la teoría normativa y la práctica política. Lo que acabo de hacer es emplear una suposición sobre las circunstancias que más favorecen la maduración de la gente, y mostrar las consecuencias para una sociedad contractualista, dada esa suposición.

Lo que quería ilustrar con mi discusión es la gran importancia que el contrato social concede al problema de la educación humana —que evidentemente refleja la importancia que la educación tiene para el florecimiento de los seres humanos—. No tendría sentido dar a los individuos oportunidades para vivir productiva y satisfactoriamente, si fueran incapaces de beneficiarse de esas oportunidades por carecer de habilidades educadas o una sensibilidad desarrollada. La idea contractualista de un acuerdo hipotético sobre las condiciones de la interacción, donde cada parte tiene en cuenta su vida entera, asegura que se dará la debida importancia a lo relativo a la crianza y educación; y no será un asunto olvidado, desatendido o infravalorado como creo que tiende a ser cuando las decisiones políticas reflejan sólo las preocupaciones presentes y futuras de personas cuya infancia, buena o mala, es ya cosa del pasado.

Pero, además, la idea contractualista del acuerdo asegura que la importancia que se da a la educación está dirigida a obligar a los futuros padres a que formen y traten de ejecutar proyectos vitales que, por así decir, internalicen los costes de sus actividades reproductivas. Como todas las grandes bendiciones, los hijos son también una carga, y si los padres disfrutaran de esas bendiciones echando sobre otros la

carga que conllevan, estarían violando el requisito fundamental del beneficio mutuo. No debemos suponer, por tanto, que la gran importancia que una sociedad contractualista concedería a la educación daría lugar a medidas que autorizaran a las personas a tener hijos y criarlos a expensas de la comunidad.

## 11

Los únicos intereses públicos y obligaciones acordes con una sociedad contractualista nacen del requisito de que todos los miembros normales de la sociedad hallen valiosa su pertenencia a la misma y aporten una contribución valiosa para los demás, conforme a condiciones igualmente aceptables para todas las personas consideradas como deliberadores racionales. Si las obligaciones ponen el énfasis en un intenso sentimiento de responsabilidad, para asegurar que cada individuo es un benefactor para sus conciudadanos, los intereses públicos hacen hincapié en que se garantice la posibilidad de que todos sean beneficiarios de las oportunidades creadas por la interacción cooperativa de esos mismos conciudadanos. Así el contractualista estaría de acuerdo con John Kennedy cuando, hacia el final de su vida, dijo: «Pregúntate lo que puedes hacer tú por tu país», pero en completo desacuerdo con el inicio de la frase: «No preguntes qué puede hacer tu país por ti». En última instancia, lo único que importa es lo que una sociedad contractualista hace por sus miembros individuales. [Sólo las oportunidades que les da justifican las obligaciones que inevitablemente impone.] Así, el contractualista es el verdadero revolucionario, que exige que realicemos una sociedad en la que cada individuo pueda florecer de acuerdo con el ideal que ha elegido para sí mismo (y no de acuerdo con ningún criterio impuesto), mientras contribuye al florecimiento de todos los demás.[16]

16. Una versión reducida de este artículo ha aparecido simultaneamente en *Proceedings of the 19th. International Wittgenstein Symposium 1996. Current Issues in Political Philosophy*, Peter Kohler y Klaus Puhl (comps.), Viena, Hoelder-Pichler-Tempsky, 1997.

# David Gauthier
# Egoísmo, moralidad y sociedad liberal

David Gauthier se convirtió, con su obra *La moral por acuerdo* (1986), en un punto de referencia del neocontractualismo liberal contemporáneo. Para este volumen, el autor canadiense ha seleccionado cuatro ensayos que acercan al lector su comprensión del contractualismo y la deliberación racional, y ha añadido un prefacio que explica la evolución y perspectiva de su proyecto. La originalidad de David Gauthier, profesor de Filosofía en la Universidad de Pittsburgh, se hace evidente en cada uno de los problemas de los que se ocupa: como teórico de la racionalidad se opone a la concepción ortodoxa de la racionalidad económica, y lo hace argumentando desde las propias premisas de ésta, en su propio terreno; como filósofo moral, intenta demostrar que el comportamiento moral es racional sin desapegarse de la visión instrumental de la racionalidad que Hume declaró impotente fuera del ámbito de los cálculos y las cuestiones de hecho; como filósofo político, considera contractualmente justificada una sociedad civil liberal, cuyas instituciones principales se articulasen en torno a la participación política y el mercado. Gauthier presenta, por tanto, una posición peculiar, que añade a estos rasgos la audacia de emplear –con un nivel de especialización notable– el mismo lenguaje técnico del modelo económico de la racionalidad cuya crítica y reformulación constituye el núcleo de su enfoque.

Pedro Francés (Barcelona, 1967) es profesor de Teoría de la Elección Racional en la Universidad Complutense e investigador visitante en el Departamento de Filosofía de la Universidad de Pittsburgh. Es autor de la tesis doctoral *El contrato moral* y de diversos artículos sobre el pensamiento de David Gauthier.

ISBN 84-493-0587-X
46054
9 788449 305870